PROCEEDINGS

IberSPEECH 2012

VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

21-23 NOVEMBER 2012
ESCUELA POLITÉCNICA SUPERIOR
UNIVERSIDAD AUTÓNOMA DE MADRID (SPAIN)

Organized by:

ATVS Biometric Recognition Group

Spanish Thematic Network on Speech Technology (RTTH)

ISCA Special Interest Group on Iberian Languages (SIG-IL)

Supported by:

Consorcio MAVIR

Proceedings published by:

Springer Communications in Computer and Information Science

PROCEEDINGS

VII Jornadas en Tecnología del Habla
and
III Iberian SLTech Workshop

November 21-23, 2012
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Editors: *Doroteo Torre Toledano*
Alfonso Ortega Giménez
António Teixeira
Joaquín González Rodríguez
Luis Hernández Gómez
Rubén San Segundo Hernández
Daniel Ramos Castro

Organized by

Supported by

ISBN: 84-616-1535-2

http://iberspeech2012.ii.uam.es/

Welcome Message

On behalf of the IberSPEECH 2012 Organizing Committee, the Spanish Thematic Network on Speech Technology (RTTH) and the ISCA-Special Interest Group on Iberian Languages (SIG-IL), it is our pleasure to welcome you to the IberSPEECH 2012 Conference, hosted by the ATVS Biometric Research Group, Universidad Autónoma de Madrid and held from 21 to 23 November 2012 in Madrid, Spain.

This conference is the result of the merging of two conferences: the "*Jornadas en Tecnología del Habla"* (Spanish Speech Technology Workshop) and the Iberian SLTech Workshop. The first one has been organized by the *"Red Temática en Tecnología del Habla"* (Spanish Speech Technology Thematic Network, http://www.rthabla.es) since 2000. This network was created in 1999 and currently includes over 200 researchers and 30 research groups in speech technology in Spain. The first Iberian SLTech Workshop was organized in Porto Salvo, Portugal, in 2009, by the Special Interest Group on Iberian Languages (SIG-IL, http://www.il-sig.org/) of the International Speech Communication Association (ISCA, http://www.isca-speech.org) and has been organized in conjunction with the *"Jornadas en Tecnología del Habla"* since 2010.

As a result, *IberSPEECH 2012: Joint "VII Jornadas en Tecnología del Habla" and III Iberian SLTech Work-shop*, is probably one of the most important research meetings in the field of speech and language processing fo-cusing on Iberian Languages, attracting many researchers (about 140 in this edition) mainly from Spain and Portu-gal, and being also a natural meeting for researchers from Latin America. However, although the main focus is on Iberian Languages and the Iberian region, the conference is not restricted to them. Proof of this are the ALBAYZIN Technology Competitive Evaluations, organized in conjunction with the conference, which in this edition have attracted the interest of several research groups from all around the world, including USA, United Kingdom, France, Japan, China, and Switzerland, among others.

The ALBAYZIN Technology Competitive Evaluations have been organized alongside with the conference since 2006, promoting the fair and transparent comparison of technology in different fields related to speech and language technology. In this edition we have had five different evaluations: Language Recognition, Audio Seg-mentation, Speech Synthesis, Search on Speech and Handwriting Recognition. The organization of each one of these evaluations requires preparing development and test data, providing data along with a clear set of rules to the participants, and gathering and comparing results from participants. This organization has been carried out by different groups of researchers and is crucial for the success of the evaluations. Results of these evaluations will be presented in a Special Session on Wednesday 21 November 2012, and are also included in the online conference proceedings.

In this edition, for the first time for this conference, Springer has accepted to publish a selection of the papers submitted to this conference in a volume of Communications in Computer and Information Science (CCIS), in particular in volume 328 entitled *Advances in Speech and Language Technologies for Iberian Languages*. Of the approximately one hundred articles accepted for presentation at the conference, only 29 of the papers accepted in the regular paper track have been selected for this publication. This selection has been based on the scores and comments provided by our Scientific Review Committee which includes over 75 researchers from different insti-tutions mainly from Spain, Portugal and Latin America, and to whom we would like to express our deepest grati-tude. Each article in the regular paper track has been reviewed by three different reviewers and authors have had some time to address the comments before submitting the camera-ready paper. The articles in the regular paper track published in Springer CCIS have been organized into six oral sessions with the following topics:

- Speaker Characterization and Recognition.
- Audio and Speech Segmentation.
- Pathology Detection and Speech Characterization.
- Dialogue and Multimodal Systems.
- Robustness in Automatic Speech Recognition.
- Applications of Speech and Language Technologies.

The other 24 articles accepted in the regular paper track have been organized into two poster sessions the first two days of the conference. Besides the regular paper track and the articles related to the ALBAYZIN Evaluations, the conference has special tracks for presenting Projects, Demos and Thesis. In this edition we have had 10 articles presenting projects, 11 articles presenting demos, and 8 articles presenting Ph.D. Thesis. All these will be present-ed in a Special Poster and Demos session the second day of the conference.

Besides the excellent research articles included in the conference, we will also have the pleasure of having three extraordinary keynote speakers: Dr. Jan "Honza" Cernocky (Brno University of Technology, BUT, Czech Repub-lic) will give the talk "Building speech recognizers with limited resources" on 21 November, Dr. Philip Rose (Australian National University, Australia) will give the talk "Not too bad? – Forensic Voice Comparison in Theo-ry and Reality" on 22 November, and finally Dr. Pedro Moreno (Google Research, NY, USA) will give the talk "Google's speech internationalization project: From 1 to 300 languages and beyond" on 23 November.

The conference is mainly organized and supported by the Spanish Thematic Network on Speech Technology *("Red Temática en Tecnología del Habla")* and the ISCA Special Interest Group on Iberian Languages (SIG-IL). Besides this, we have also received support from the Universidad Autónoma de Madrid (UAM) and the Campus Internacional Excelencia UAM+CSIC, which have not only provided a fantastic venue for organizing the confer-ence (the Escuela Politécnica Superior), but also financial support. Also, several companies have provided finan-cial support for the conference, including Google, Microsoft and Telefónica (through Cátedra UAM-Telefónica). Last but not least, we have had financial support from the MA2VICMR consortium. Without the financial support of all of them this conference would simply have not been possible.

We would also like to thank Springer, and in particular to Alfred Hoffmann and Leonie Kunz, for the possibility of publishing a selection of the articles in this conference in a volume of Communications in Computer and Information Science and their help and great work in preparing it. This will help increasing the international impact of this conference.

Finally we would like to thank all the people that have been putting the best or their efforts in making this con-ference a reality: the Organizing Committee, the Local Organizing Committee, the organizations and companies that have provided financial support, the reviewers, the authors, the organizers and participants in the ALBAYZIN evaluations. We join all of them in wishing all attendees a fruitful and very enjoyable conference.

Doroteo Torre Toledano
Alfonso Ortega Giménez
António Teixeira

IberSPEECH 2012 Co-Chairs

Organization

General Chairs

Doroteo Torre Toledano	Universidad Autónoma de Madrid, Spain
Alfonso Ortega Giménez	Universidad de Zaragoza, Spain
António Teixeira	Universidade de Aveiro, Portugal

Program Chairs

Joaquín González Rodríguez	Universidad Autónoma de Madrid
Rubén San Segundo Hernández	Universidad Politécnica de Madrid
Luis Hernández Gómez	Universidad Politécnica de Madrid

Publication Chair

Daniel Ramos Castro	Universidad Autónoma de Madrid

Demos Chairs

Daniela Braga	Microsoft (MLDC)
Alberto Abad Gareta	INESC

Awards Chairs

Miguel Sales Dias	Microsoft (MLDC)
Climent Nadeu Camprubi	Universitat Politècnica de Catalunya

Evaluation Chairs

Javier Tejedor Noguerales	Universidad Autónoma de Madrid
Javier González Domínguez	Universidad Autónoma de Madrid

Local Organizing Committee

Doroteo Torre Toledano	Universidad Autónoma de Madrid
Joaquín González Rodríguez	Universidad Autónoma de Madrid
Daniel Ramos Castro	Universidad Autónoma de Madrid
Javier González Domínguez	Universidad Autónoma de Madrid
Javier Franco Pedroso	Universidad Autónoma de Madrid
Javier Tejedor Noguerales	Universidad Autónoma de Madrid
Fernando Jesús López Colino	Universidad Autónoma de Madrid
Javier Ortega García	Universidad Autónoma de Madrid
Julián Fiérrez Aguilar	Universidad Autónoma de Madrid
Javier Galbally Herrero	Universidad Autónoma de Madrid
Rubén Vera Rodríguez	Universidad Autónoma de Madrid
Pedro Tomé González	Universidad Autónoma de Madrid
Álvaro Diéguez Sánchez-Largo	Universidad Autónoma de Madrid
Alicia Lozano Díez	Universidad Autónoma de Madrid
Fernando Espinoza Cuadros	Universidad Autónoma de Madrid
Rubén Zazo Candil	Universidad Autónoma de Madrid
Leonardo Campillo Llanos	Universidad Autónoma de Madrid
Olga León Zurdo	Universidad Autónoma de Madrid
Antonio Moreno Sandoval	Universidad Autónoma de Madrid
Sara Antequera Sanz	Universidad Autónoma de Madrid

Scientific Review Committee

Alberto Abad Gareta	INESC
Olatz Arregi Uriarte	Euskal Herriko Unibertsitatea
José Miguel Benedí Ruiz	Universitat Politécnica de Valencia
Mª Carmen Benítez Ortúzar	Universidad de Granada
Antonio Bonafonte Cávez	Universitat Politècnica de Catalunya
Daniela Braga	MLDC/Microsoft
Francisco Campillo	Universidad de Vigo
Antonio Cardenal	Universidade de Vigo
Valentín Cardeñoso Payo	Universidad de Valladolid
Paula Carvalho	Universidade de Lisboa
Francisco Casacuberta Nolla	Universitat Politécnica de Valencia
María José Castro Bleda	Universitat Politécnica de Valencia
José Colás Pasamontes	Universidad Autónoma de Madrid
Ricardo de Córdoba Herralde	Universidad Politécnica de Madrid
Carmen de la Mota Gorriz	Universitat Autònoma de Barcelona
Ángel de la Torre Vega	Universidad de Granada
Laura Docío Fernández	Universidade de Vigo
Daniel Erro Eslava	Euskal Herriko Unibertsitatea
David Escudero Mancebo	Universidad de Valladolid
Rubén Fernández Pozo	Universidad Politécnica de Madrid
Javier Ferreiros López	Universidad Politécnica de Madrid
Ascensión Gallardo Antolín	Universidad Carlos III de Madrid
Fernando García Granada	Universitat Politécnica de Valencia
Carmen García Mateo	Universidade de Vigo
Marta Gatius Vila	Universitat Politècnica de Catalunya
Juan Ignacio Godino Llorente	Universidad Politécnica de Madrid
Koldo Gojenola Galletebeitia	Euskal Herriko Unibertsitatea
Pedro Gómez Vilda	Universidad Politécnica de Madrid
Javier González Domínguez	Universidad Autónoma de Madrid
Joaquín González Rodríguez	Universidad Autónoma de Madrid
Inma Hernaez Rioja	Euskal Herriko Unibertsitatea
Luis A. Hernández Gómez	Universidad Politécnica de Madrid
Fco. Javier Hernando Pericas	Universitat Politècnica de Catalunya
Lluís Felip Hurtado Oliver	Universitat Politécnica de Valencia
Eduardo Lleida Solano	Universidad de Zaragoza
Joaquím Llisterri Boix	Universitat Autònoma de Barcelona
Fernando Jesús López Colino	Universidad Autónoma de Madrid
Eduardo López Gonzalo	Universidad Politécnica de Madrid
María Teresa López Soto	Universidad de Sevilla
Ramón López-Cózar Delgado	Universidad de Granada
Nuno Mamede	INESC
José B. Mariño Acebal	Universitat Politècnica de Catalunya
Carlos David Martínez Hinarejos	Universitat Politécnica de Valencia
Helena Moniz	INESC
Juan Manuel Montero Martínez	Universidad Politécnica de Madrid
Nicolás Morales Mombiela	Nuance
Asunción Moreno Bilbao	Universitat Politècnica de Catalunya
Climent Nadeu Camprubi	Universitat Politècnica de Catalunya
Juan Luis Navarro Mesa	Universidad de Las Palmas de Gran Canaria
Eva Navas Cordón	Euskal Herriko Unibertsitatea
Juan Nolazco Flores	Tecnológico de Monterrey
Alfonso Ortega Giménez	Universidad de Zaragoza
Antonio Miguel Peinado Herreros	Universidad de Granada

Carmen Peláez Moreno	Universidad Carlos III de Madrid
José Luis Pérez Córdoba	Universidad de Granada
Paulo Quaresma	Universidade de Évora
Daniel Ramos Castro	Universidad Autónoma de Madrid
Andreia Rauber	Universidade Católica de Pelotas
José Adrián Rodríguez Fonollosa	Universitat Politècnica de Catalunya
Eduardo Rodríquez Banga	Universidade de Vigo
Luis Javier Rodríguez Fuentes	Euskal Herriko Unibertsitatea
Antonio José Rubio Ayuso	Universidad de Granada
Victoria Eugenia Sánchez Calle	Universidad de Granada
Joan Andreu Sánchez Peiró	Universitat Politécnica de Valencia
Emilio Sanchís Arnal	Universitat Politécnica de Valencia
Rubén San-Segundo Hernández	Universidad Politécnica de Madrid
Kepa Sarasola Gabiola	Euskal Herriko Unibertsitatea
Ibon Saratxaga Couceiro	Euskal Herriko Unibertsitatea
Encarnación Segarra Soriano	Universitat Politécnica de Valencia
Mário Silva	Universidade de Lisboa
Alberto Simões	ESEIG/IPP
Daniel Tapias Merino	SIGMA Technologies
António Teixeira	Universidade de Aveiro
Javier Tejedor Noguerales	Universidad Autónoma de Madrid
Doroteo Torre Toledano	Universidad Autónoma de Madrid
María Amparo Varona Fernández	Euskal Herriko Unibertsitatea
José Luis Vicedo González	Universitat d'Alacant

Conference Program at a Glance

Wednesday 21 November	Thursday 22 November	Friday 23 November
8:00 On-site registration		
8:30 Opening Ceremony	8:45 On-site registration	8:45 On-site registration
9:00 O1 – Oral Session *Speaker Recognition and Characterization*	9:00 O3 – Oral Session *Pathology and Speech Characterization*	9:00 O4 – Oral Session *Robustness in Automatic Speech Recognition*
10:40 Coffee Break	10:40 Coffee Break	10:40 Coffee Break
11:00 **Keynote Talk** ***Jan "Honza" Cernocky***	11:00 **Keynote Talk** ***Philip Rose***	11:00 **Keynote Talk** ***Pedro Moreno***
12:00 P1 – Poster Session	12:00 P2 – Poster Session	12:00 O6: Oral Session *Application of Speech & Language Technologies*
13:30 – Lunch	13:30 – Lunch	13:30 – Closing Ceremony
15:00 O2 – Oral Session *Audio and Speech Segmentation*	15:00 O4 – Oral Session *Dialogue and Multimodal Systems*	
16:40 Coffee Break	16:40 Coffee Break	
17:00 Albayzin Evaluations	17:00 Projects, Demos and Thesis Session	
18:30 RTTH Assembly	18:30 Guided Visit to Museo Reina Sofía	
19:30 Welcome Reception		
	21:00 Gala Dinner at Restaurant Samarkanda	

Table of Contents

Keynote Talks..... 21

Building Speech Recognizers with Limited Resources.....23
Jan "Honza" Cernocky

Where the Science Ends and the Law Begins: Theory and Reality in Likelihood Ratio‐based Forensic Voice Comparison 25
Philip Rose

Google's Speech Internationalization Project: From 1 to 300 Languages and Beyond 41
Pedro Moreno

Oral Session O1 – Speaker Characterization and Recognition
All papers in this sesión are included in Springer CCIS vol. 328

Reliability Estimation of the Speaker Verification Decisions Using Bayesian Networks to Combine Information from Multiple Speech Quality Measures
Jesus Villalba, Eduardo Lleida Solano, Alfonso Ortega Giménez, and Antonio Miguel Artiaga

On the use of Total Variability and Probabilistic Linear Discriminant Analysis for Speaker Verification on Short Utterances
Javier González Domínguez, Rubén Zazo, and Joaquin González-Rodríguez

Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition
Javier Franco Pedroso, Fernando Espinoza, and Joaquin González-Rodríguez

Improving the Quality of Standard GMM-based Voice Conversion Systems by Considering Physically Motivated Linear Transformations
Tudor-Cătălin Zorilă, Daniel Erro Eslava, and Inma Hernaez Rioja

Evaluation of a New Beam-Search Formant Tracking Algorithm in Noisy Environments
Dayana Ribas Gonzalez, Jose Enrique Garcia Lainez, Antonio Miguel Artiaga, Alfonso Ortega Giménez, Eduardo Lleida Solano, and Jose Ramon Calvo de Lara

Poster Session P1..... 43

Audio Encoding for Heart and Breath Sounds Acquired with Digital Stethoscope.....45
Ignacio Palacios Santos, Doroteo Torre Toledano, and Andrés Martínez Fernández

Morphological Processing of a Dynamic Compressive Gammachirp Filterbank for Automatic Speech Recognition 53
Joyner Cadore, Carmen Peláez-Moreno, and Ascensión Gallardo Antolín

Manipulation of Unvoiced Fricatives for Improved Discrimination by Hearing Impaired 63
Rami Cohen, Oren Ierushalmi, Yizhar Lavner, Michal Genussov, and Ami Steiner

Multiclass Pitch Accent Classification for Assisting Manual Prosodic Labeling.....73
David Escudero Mancebo , Francisco Vicaíno-Ortega, César González-Ferreras, Carlos Vivaracho-Pascual, Mercedes Cabrera-Abreu, Eva Estebas-Vilaplana, and Valentín Cardeñoso Payo

Language Identification based on a Discriminative Text Categorization Technique 83
Miguel Caraballo, Ricardo de Cordoba Herralde, Luis D'Haro, Rubén San-Segundo and José Pardo

Binary Position Assignment of Two Known Simultaneous Acoustic Sources.....93
Rupayan Chakraborty, Climent Nadeu Camprubi, and Taras Butko

Emotions recognition using binary fingerprints.....103
Xavier Anguera, Esperança Movellan, and Miquel Ferrarons

Measuring Acoustic Reduction in Feature Space.....113
Henrik Schulz, and José Adrián Rodríguez Fonollosa

A preliminary approach to forensic speaker recognition using phonemes.....123
Pedro Univaso, Miguel Martínez Soler, Diego Evin, and Jorge Gurlekian

Spectral Envelope Representation using Sums of Gaussians.....133
Anderson Machado, Antonio Bonafonte Cávez, and Marcelo Queiroz

Towards a Robust Dynamic Frequency Warping Text-Independent Voice Conversion System 143
Tudor-Cătălin Zorilă, Daniel Erro, Yannis Stylianou, and Inma Hernáez

Oral Session O2 – Audio and Speech Segmentation

All papers in this sesión are included in Springer CCIS vol. 328

On the Influence of Automatic Segmentation and Clustering in Automatic Speech Recognition
Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo, and Antonio Cardenal-Lopez

Preliminary Results of Alignment of Text and Audio in News and Song
Darwin Patricio Córdova Lucero and Doroteo Torre Toledano

Aligning very long speech signals to bilingual transcriptions of parliamentary sessions
German Bordel, Mikel Penagarikano, Luis Javier Rodríguez Fuentes, and María Amparo Varona Fernández

Factor Analysis Segmentation and Classification in Broadcast News Domain
Diego Castán, Alfonso Ortega Giménez, and Eduardo Lleida Solano

Prosodic and Phonetic Features for Speaking Styles Classification and Detection
Arlindo Veiga, Dirce Celorico, Jorge Proença, Sara Candeias, and Fernando Perdigão

Oral Session O3 – Pathology Detection and Speech Characterization

All papers in this sesión are included in Springer CCIS vol. 328

Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using Mul-tiFocal Toolkit
David Martínez González, Eduardo Lleida Solano, Alfonso Ortega Giménez, Antonio Miguel Artiaga, and Jesus Villalba

Score Level versus Audio Level Fusion for Voice Pathology Detection on the Saarbrücken Voice Database
David Martínez González, Eduardo Lleida Solano, Alfonso Ortega Giménez, and Antonio Miguel Artiaga

Using HMM to detect speakers with severe Obstructive Sleep Apnoea Syndrome
Ana Montero Benavides, José Luis Blanco, Alejandra Fernández, Rubén Fernandez Pozo, Doroteo Torre To-ledano, and Luis Hernández Gómez

Acoustic analysis of European Portuguese oral vowels produced by children
Catarina Oliveira, Maria Manuel Cunha, Samuel Silva, António Teixeira, and Pedro Couto

Impact of age in ASR for the elderly: preliminary experiments in European Portuguese
Thomas Pellegrini, Isabel Trancoso, Annika Hamalainen, António Calado, Miguel Dias, and Daniela Braga

Poster Session P2..... 153

Rule-Based Pronunciation Modeling for Resource-Scarce Varieties.....155
Simone Ashby, José Pedro Ferreira, Sílvia Barbosa, and Catarina Silva

A quantitative study of disfluencies in formal, informal and media spontaneous speech in Spanish 164
Antonio Moreno Sandoval, Leonardo Campillos Llanos, and Doroteo Torre Toledano

A Computational Morphology for Spanish Sign Language Nominal Inflection 173
Jordi Porta, Fernando Jesús López Colino, and José Colás Pasamontes

Modeling a Spoken Dialog System using POMDPs and Agenda-based User Simulation 183
David Griol, José Manuel Molina, Araceli Sanchis de Miguel, and Zoraida Callejas

Methodology for developing a Speech into Sign Language Translation System in a New Semantic Domain 193
Verónica López-Ludeña, Rubén San-Segundo, Carlos González, Juan López, and José Pardo

Architecture for Text Normalization using Statistical Machine Translation techniques 204
Verónica López-Ludeña, Rubén San-Segundo, Juan Manuel Montero, Roberto Barra-Chicote, and Jaime Lorenzo

GenProso: a parametric prosody prediction module for text-to-speech applications.....214
Juan María Garrido

MAVIR: a corpus of spontaneous formal speech in Spanish and English 224
Leonardo Campillos Llanos and Antonio Moreno Sandoval

Velar Movement in European Portuguese Nasal Vowels 231
Paula Martins, Catarina Oliveira, Samuel Silva, and António Teixeira

Subjective Evaluation of a High Quality Text-to-Speech System for Argentine Spanish 241
Jorge Gurlekian, Christian Cossio Mercado, Humberto Torres and M. Elena Vaccari

Design of a message verification tool to be implemented in CALL systems 251
Igor Odriozola, Inma Hernáez, and Eva Navas

Vocalic Filler Characterization using Acoustical Properties Derived from Cepstrum 260
Arlindo Veiga, Jorge Proença, Dirce Celorico, Sara Candeias and Fernando Perdigão

Oral Session O4 – Dialogue and Multimodal Systems
All papers in this sesión are included in Springer CCIS vol. 328

Mutual Information and Perplexity based Clustering of Dialogue Information for Dynamic Adaptation of Lan-guage Models
Juan Lucas-Cuesta, Fernando Fernández, Tirso Moreno, and Javier Ferreiros López

A multilingual SLU system based on semantic decoding of graphs of words
Marcos Calvo, Lluís Hurtado, Fernando García Granada, and Emilio Sanchís Arnal

Merging intention and emotion to develop adaptive dialogue systems
Zoraida Callejas, David Griol and Ramón López-Cózar Delgado

Language Technology for Handwritten text Recognition
Alejandro Toselli, Nicolas Serrano, Adrià Giménez-Pastor, Ihab Khoury, Alfons Juan, and Enrique Vidal

Character-based Handwritten Text Recognition of Multilingual Documents
Miguel del Agua, Nicolás Serrano Martínez-Santos, Jorge Civera, and Alfons Juan

Projects, Demos and Thesis Special Session – Projects 269

Architecture for Text Sign Localization and Recognition 271
Arturo Romero, José Moreno, Fernando Fernández, Juan Manuel Montero and Rubén San-Segundo

TexAFon: a multilingual text processing tool for text-to-speech applications 281
Juan María Garrido, Sílvia Rustullet, Yesika Laplaza, Montserrat Marquina, and Claudia Schoenfelder

The First European Portuguese Elderly Speech Corpus 290
Annika Hamalainen, Fernando Miguel Pinto, Miguel Dias, Ana Júdice, João Freitas, Carlos Pires, Vitor Duarte Teixeira, António Calado, and Daniela Braga

Let's Speak Quechua: The Implementation of a Text-to-Speech System for the Incas' Language 300
Richard Alexander Castro Mamani, John Edgar Vargas Muñoz, and Juan Cruz

Cotovía: an open source TTS for Galician and Spanish, Eduardo Rodríguez Banga 308
Carmen Garcia-Mateo, Francisco Méndez Pazó, Manuel González González, and Carmen Magariños Iglesias

The Glissando project: Building and exploitation of a large bilingual annotated corpus for multidisciplinary prosodic analysis and applications 316
Juan María Garrido, David Escudero Mancebo, and Lourdes Aguilar.

The CASMACAT Project: The Next Generation Translator's Workbench 326
Daniel Ortiz-Martínez, Germán Sanchis-Trilles, Francisco Casacuberta Nolla, Vicent Alabau, Enrique Vidal, José-miguel Benedí, Jesús González-Rubio, Alberto Sanchís, and Jorge González

Google's Voice Search in Iberian Languages, Ignacio Lopez Moreno 335
Fadi Biadsy, Eugene Weinstein, and Pedro Moreno.

transLectures 345
Alfons Juan, Joan Albert Silvestre, Miguel del Agua, Gonçal Garcés, Guillem Gascó, Adrià Giménez-Pastor, Adrià Martínez, Alejandro Pérez, Isaías Sánchez, Nicolás Serrano Martínez-Santos, Rachel Spencer, Juan Daniel Valor Miró, Jesús Andrés-Ferrer, Jorge Civera, and Alberto Sanchís

META-NET. Official Languages of Spain in the Digital Age 352
Asunción Moreno Bilbao, Núria Bel, Maite Melero, Carmen García-Mateo, Inma Hernaez Rioja, Sergio Oller, aljoscha Burchardt, Kathrin Eichler, Georg Rehm, and Hans Uszkoreit

Projects, Demos and Thesis Special Session – Demos 363

Hermes: a mobile distributed speech translation system 365
Emilio Granell, Carlos David Martínez Hinarejos, and Vicent Tamarit

COPS: a computer programming tool to cope with functional diversity 371
Carlos David Martínez Hinarejos, Santiago Sánchez-Alepuz and Natividad Prieto-Sáez

GTMTranscriber: a Subtitling Program for Broadcast News in Galician Language 377
Antonio Cardenal-Lopez, Carmen Garcia-Mateo, Alberto Espiña, Paula Lopez-Otero, and Laura Docio-Fernandez

An on-line system adding subtitles and sign language to Spanish audio-visual content 383
Jordi Porta, Fernando Jesús López Colino, Javier Tejedor, and José Colás Pasamontes

EntrenaVoz. Videogame feedback system for speech training and rehabilitation 389
Victor Osma-Ruiz, Nicolás Sáenz-Lechón, Juana Gutiérrez-Arriola, Ruben Fraile, and Juan Ignacio Godino Llorente

Interactive and User-Adapted Voice Portals for the City Council 395
David Griol, María García Jiménez, José Manuel Molina, and Araceli Sanchis de Miguel

The MovieScript System: Integration of Statistical Methodologies and the VoiceXML standard to develop spoken dialog systems 400
David Griol, José Manuel Molina, and Araceli Sanchis de Miguel

VITHEA: On-line word naming therapy in Portuguese for aphasic patients exploiting automatic speech recognition 405
Anna Pompili, Pedro Fialho, and Alberto Abad

Learning Portuguese with Speech Technologies 411
Isabel Trancoso, Thomas Pellegrini, André Silva, Rui Correia, Nuno Mamede and Jorge Baptista

Aromo: Argentine Spanish TTS System 416
Humberto Torres, Jorge Gurlekian, and Christian Cossio Mercado

Search Engine for Multilingual Audiovisual Contents 422
José David Pérez, Antonio Bonafonte Cávez, Antonio Cardenal, Marta Ruiz Costa-jussa, José Adrián Rodríguez Fonollosa, Asunción Moreno Bilbao, Eva Navas Cordón, and Eduardo Rodríguez Banga

Projects, Demos and Thesis Special Session – Ph.D. Thesis 431

Application of Speech Technology in the Education of Children's Altered Voice 433
William Ricardo Rodríguez Dueñas

Advances in Fully-Automatic and Interactive Phrase-based Statistical Machine Translation. 437
Daniel Ortiz-Martínez

Session Variability Compensation in Automatic Speaker and Language Recognition 447
Javier González Domínguez

Detection and Handling of Overlapping Speech for Speaker Diarization 460
Martin Zelenák

Phase in the harmonic models of the speech signal: strategies for representation, processing and applications 470
Ibon Saratxaga Couceiro

Building task-oriented machine translation systems 480
Germán Sanchis-Trilles

Robust Diarization for Speaker Characterization 490
Carlos Vaquero

Perceptual Optimization of Unit-Selection Text-to-Speech Synthesis Systems by means of active interactive Genetic Algorithms 500
Lluís Formiga

Oral Session O5 – Robustness in Automatic Speech Recognition..... 511
All papers in this sesión except the last one are included in Springer CCIS vol. 328

A robust pitch extractor based on DTW lines and CASA with application in noisy speech recognition
Juan Morales Cordovilla, Pablo Cabañas-Molero, Antonio Miguel Peinado Herreros, and Victoria Eugenia Sánchez Calle

Speech Denoising Using Non-Negative Matrix Factorization with Kullback-Leibler Divergence and Sparseness Constraints
Jimmy Ludeña Choez and Ascensión Gallardo Antolín

MMSE feature reconstruction based on an occlusion model for robust ASR
Jose Gonzalez, Antonio Miguel Peinado Herreros, and Angel Gomez

Automatic Speech Recognition based on Ultrasonic Doppler Sensing for European Portuguese
João Freitas, António Teixeira, Francisco Vaz, and Miguel Dias

Pairwise Likelihood Normalization-Based Channel Selection for Multi-Microphone ASR ...513
Martin Wolf and Climent Nadeu Camprubi

Oral Session O6 – Applications of Speech and Language Technologies
All papers in this sesión are included in Springer CCIS vol. 328

Integrating a state-of-the-art ASR system into the Opencast Matterhorn platform
Juan Daniel Valor Miró, Alejandro Pérez González de Martos, Jorge Civera, and Alfons Juan

Speech Reconstruction by Sparse Linear Prediction
Jan Koloda, Antonio Miguel Peinado Herreros, and Victoria Eugenia Sánchez Calle

Steganographic pulse-based recovery for robust ACELP transmission over erasure channels
Domingo López, Angel Gomez, José Luis Pérez Córdoba, Bernd Geiser, and Peter Vary

A proposal for a visual speech animation system for European Portuguese
José Serra, Manuel Ribeiro, João Freitas, Verónica Orvalho, and Miguel Dias

Online learning of log-linear weights in interactive machine translation
Francisco Javier López-Salcedo, Germán Sanchis-Trilles, and Francisco Casacuberta Nolla

Albayzin Language Recognition Evaluation System Descriptions 523

ANHYT 525
Fusheng Ba, Xuefei Jia, Zhihua Jin,Rui Xing, and Xu Ji

GTH 528
Luis Fernando D'Haro and Ricardo Córdoba

L^2F 540
Alberto Abad

MSMP-LaBRI 550
Jean Luc Rouas

THU-EE 554
Jia Liu, Liang He, Jiaming Xu, Weiwei Liu, and Wei-Qiang Zhang

UVigo-GTM 557
Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

ViVoLab UZ 566
David Martínez, Eduardo Lleida, and Alfonso Ortega

Albayzin Audio Segmentation Evaluation System Descriptions 575

Aholab 577
David Tavarez, Eva Navas, Daniel Erro, and Ibon Saratxaga

CAIAC-UAM 585
Héctor Delgado and Javier Serrano

GTTS 590
Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez and Germán Bordel

PRHLT-UPV 596
J. A. Silvestre-Cerdá, A. Giménez, J. Andrés-Ferrer, J. Civera and A. Juan

UVigo-GTM 601
Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Albayzin Search-on-Speech Evaluation System Descriptions 609

ELiRF 610
Emilio Sanchis, Lluís-F. Hurtado, Jon A. Gómez, Marcos Calvo and Raúl Fabra

GTTS 619
Amparo Varona, Mikel Penagarikano, Luis Javier Rodríguez-Fuentes, Germán Bordel and Mireia Diez

Telefonica Research 626
Xavier Anguera

ViVoLab UZ 633
Anotnio Miguel, Jesus Villalba, Eduardo Lleida and Alfonso Ortega

Albayzin Speech Synthesis Evaluation System Descriptions 643

Aholab 645
David Tavarez, Eva Navas, Daniel Erro, and Ibon Saratxaga

Simple4All 653
Jaime Lorenzo-Trueba, Oliver Watts, Roberto Barra-Chicote, Junichi Yamagishi, Simon King and Juan M. Montero

Keynote Talks

Keynote talk

Title: Building speech recognizers with limited resources

Speaker: Jan Honza Černocký

With contributions of: Martin Karafiát, Miloš Janda, Mirko Hannemann, Karel Veselý, František Grézl, Ekaterina Egorova

Abstract:

Research groups and industry have reached excellent results in automatic speech recognition (ASR) of language with abundant speech and text resources. There is however a steady need for recognition, keyword spotting and spoken term detection systems in languages that lack these resources.

This talk will summarize the work Brno University of Technology (BUT) speech group has done so far in this field. Supported by a project of Czech Ministry of Trade and Commerce and by U.S. IARPA BABEL program, we have investigated approaches that allow for re-using data from well represented languages to obtain better results in the recognition of low-resource ones: in feature extraction, we have worked on language independency of nowadays very popular bottle-neck features generated by neural networks. We have also studies sub-space Gaussian Mixture Models (SGMMs) and Regional-Dependent Linear Transforms (RDLT) for acoustic modeling and adaptation, respectively. Finally, we will present several ideas in automatic building of pronunciation dictionaries.

Speaker Bio:

Jan "Honza" Cernocky http://www.fit.vutbr.cz/~cernocky (Ing. [MS] 1993 Brno University of Technology (BUT); Dr. [PhD] 1998 Universite Paris XI and BUT) was with the Institute of Radio-electronics, BUT (Faculty of Electrical Engineering and Computer Science) as assistant professor from 1997. Since February 2002, he is with the Faculty of Information Technology (FIT), BUT as Associate Professor (Doc.) and Head of the Department of Computer Graphics and Multimedia. With Dr. Burget and Prof. Hermansky, he is leading the BUT Speech@FIT research group. Dr.

Cernocky supervises several PhD students. He has been involved with several European projects: SPEECHDAT-E (4th FP, technical coordination), SpeeCon, Multimodal meeting manager (M4, both 5th FP), Augmented Multiparty interaction (AMI, 6th FP), Augmented Multiparty Interaction with Distance Access (AMIDA, 6th FP), and CareTaker (6th FP). He was BUT's principal investigator in EC-sponsored MOBIO project (7th FP) and in projects supported by Czech Ministries of Industry and Interior. He is currently coordinating a project supported by Technology Agency of the Czech Republic (TACR) and supervises projects funded by U.S. Government.

His research interests include signal processing and speech processing (speaker and language recognition, keyword spotting and spoken term detection). He is author or co-author of more than 50 papers in journals and at conferences. He has served as reviewer for conferences and journals,

including IEEE Transactions on Speech, Audio and Language Processing. He is on the scientific board of FIT, scientific board of Text-Speech-Dialogue conference, editorial board of the journal Radioengineering, on the board of Czechoslovak section of IEEE, and on the IEEE SLTC technical committee. He served as general co-chair of ICASSP 2011 in Prague.

As faculty member of FIT BUT, Dr. Cernocky is also active in teaching, he is responsible for signal-processing, pattern recognition, speech and natural language processing courses in all levels of studies (bachelor, master, doctoral).

Dr. Cernocky is a senior member of IEEE and member of ISCA.

Keynote talk

Title: Not too bad? – Forensic Voice Comparison in Theory and Reality.

Abstract:

In forensic voice comparison, speech recordings from an unknown voice, usually of an offender, are compared with recordings from a known voice, usually the suspect. The aim, of course, is to help the trier-of-fact decide whether the suspect has said the incriminating speech.

For some time now, the use of a likelihood ratio has been both theoretically recognised as the correct logical framework for the evaluation of forensic evidence, and implemented as a matter of course in some areas, e.g. DNA profiling. It is logically correct, since by Bayes' Theorem a posterior probability – like "it is highly likely the suspect said the incriminating speech" – cannot be estimated absent prior odds, to which the expert is not usually privy. Since a posterior may well impinge on considerations of ultimate issue, the use of a likelihood ratio may also be the legally correct option.

The use of likelihood ratios in forensic voice comparison was an idea whose time came around the beginning of the new century, when its efficacy first began to be demonstrated both with automatic and traditional phonetic features (Morrison 2009a). The results from now well over a decade's extensive, and continuing, research testing have shown that the approach works rather well – same-speaker speech samples can for example be rather well discriminated from different-speaker speech samples on the basis of their LRs (although in real cases discrimination is not the appropriate model) – and it has been shown that the approach can emulate the DNA gold-standard (Gonzalez-Rodriguez et al. 2007).

However, despite these results, and the many explanatory texts written for the legal profession (e.g. Robertson & Vignaux 1995, Aitken et al. n.d.), it is safe to say the idea has not yet caught on. Although forensic voice comparison is experiencing a remarkable paradigm shift in theory, in reality this is only in a few small areas in the world (appropriately for this conference, Spain and Australia); and even then only sporadically and in conditions where it is not at all clear that the interested parties actually understand the concepts. This is partly because of the inherent conservatism of the legal profession, and partly because, it appears, many of its practitioners, if they are aware of the approach at all, find it difficult to understand. For example, in a recent set of Australian draft standards, representatives of organizations including ***The Australian Attorney-General's Department, The Australian and New Zealand Forensic Science Society, The University of New South Wales' Expertise, Evidence and Law Program, and The New South Wales Bar Association*** maintained that the "Interpretation [of forensic evidence] includes answering the question as to whether or not … items share a common origin … ." (Standards Australia 2012). Not just the legal profession finds it difficult, however. There have been many objections to the implementation of the approach – some well-founded and some simply misconceived – among many, mostlyUK, practitioners (Morrison 2009b). And perhaps the best-known phonetics text-book (Ladefoged 2006) defines the likelihood ratio as "… the likelihood that the two voices in question are the same as compared with the likelihood that they are different." thus confusing it with the prior odds.

In my talk, I want to first rehearse the conceptual background to likelihood ratio-based forensic voice comparison, and say what I see as the two immense barriers to the acceptance of the

approach where it matters – in the courtroom. I then describe a real case where likelihood ratio-based forensic voice comparison played a role in the successful prosecution of a $150 million telephone fraud. This will give an idea of what the approach actually involves. The case I describe is important for several reasons unrelated to the size of the sum of money involved. It remains the only one where likelihood ratio-based forensic voice comparison has been received in an Australian court, and is a nice example of where traditional phonetic features, as opposed to automatic features, can come into their own. Finally it is important because it suggests that at least one of my barriers – the impasse at the boundary between likelihood ratio and its matrix concept of Bayes' Theorem – may, in the reality of the courtroom as opposed to the theory of our research laboratories, be imaginary.

Aitken, C.G.G., Roberts, P., & Jackson, G. no date ***Fundamentals of Probability and Statistical Evidence in Criminal Proceedings – Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses***, Royal Statistical Society.

Gonzalez-Rodriguez J., Rose P., Ramos, D., Torre, D. & Ortega-Garcia, J. 2007 "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", ***IEEE Trans. on Audio Speech and Language Processing***, 15(7): 2104 – 2115.

Ladefoged, P. 2006 ***A Course in Phonetics***, 5th ed. Thomson.

Morrison, G.S. 2009a "Forensic voice comparison and the paradigm shift", ***Science & Justice***, 49: 298–308.

Morrison, G.S. 2009b "Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework", ***Australian Journal of Forensic Sciences***, 41: 155-161.

Robertson, B., & Vignaux, T. 1995 ***Interpreting Evidence***, Wiley.

Standards Australia 2012 "Forensic Analysis Part 3: Interpretation", Draft for Public Comment, DR AS 5388.3 http://infostore.saiglobal.com/store/Details.aspx?ProductID=1530224.

Speaker Bio:

Phil Rose is adjunct associate professor in the School of Language Studies at the Australian National University, where he taught Phonetics and Chinese Linguistics for thirty years. He studies phonetics and phonology, and is an expert on tone languages and on Chinese and Chinese dialects. He is also an expert on forensic voice comparison and in his 2002 and 2003 books pioneered the application of the Likelihood Ratio of Bayes' Theorem to traditional forensic voice comparison, thus bringing it in line with forensic DNA profiling. He has done forensic voice comparison case-work on Australian English and varieties of Chinese for nearly 20 years. He has been a British Academy visiting professor at the University of Edinburgh's ***Joseph Bell Centre for Forensic Statistics and Legal Reasoning***, is chairman of the Forensic Speech Science Committee of the ***Australasian Speech Science & Technology Association*** and a member of the ***Australian Academy of Forensic Sciences***. His attempt to retire from teaching in 2009 failed and in the first half of 2012 he was visiting professor at the Hong Kong University of Science and Technology, where he taught courses on ***Chinese Phonetics*** and ***Forensic Voice Comparison in Cantonese***.

Where the Science Ends and the Law Begins: Theory and Reality in Likelihood Ratio-based Forensic Voice Comparison.

Phil Rose

Hong Kong University of Science & Technology and Australian National University
philip.rose@anu.edu.au

Abstract. The first use of Likelihood Ratios for the evaluation of acoustic-phonetic forensic voice comparison evidence in a real trial in Australia is described and critiqued.

Keywords. Forensic voice comparison, likelihood ratio, F-pattern, fundamental frequency.

1 Introduction

In forensic voice comparison (FVC), speech recordings from an unknown voice, usually of an offender, are compared with recordings from a known voice, usually the suspect. The aim is to help the trier-of-fact (in the case described here, a 12 person jury) decide whether the suspect has said the incriminating speech. Currently FVC, as reflected in research and practice, can be divided into two types depending on what the expert considers to be their ultimate purpose. In the first type, the expert considers their aim to be to say how likely it is, given the evidence, that the suspect said the incriminating speech. For example in a 2004 high profile murder case one of the leading UK forensic speech scientists was described as saying "I came to the view that there were very strong similarities and no differences of a kind that could cause me to eliminate [the suspect]." and "I would consider the likelihood of the voices coming from two different people to be remote." [1]. In the second type of FVC, the expert's aim is seen as restricted to estimating the strength of the speech evidence with a Likelihood Ratio (LR) – in other words, to estimate how much more likely the difference between the suspect and offender speech samples is, assuming the offender sample has come from the suspect, rather than from another randomly chosen speaker in the relevant population.

For some time now, the use of a LR has been theoretically recognised as the correct logical framework for the evaluation of forensic evidence. It is referred to in all major text-books on forensic statistics, e.g [2 – 3], and has been endorsed by several forensic institutions, e.g the UK and Irish Association of Forensic Science Providers has made it part of its *Standards for the formulation of evaluative forensic science*

expert opinion [4]. It is logically correct, since by Bayes' Theorem a posterior probability – like "it is highly likely the suspect said the incriminating speech" – cannot be estimated absent prior odds, to which the expert is not usually privy. This is why the first type of FVC, unless the prior odds are known to the expert, cannot be logically correct. Since a posterior may well impinge on considerations of ultimate issue, which is up to the trier-of-fact to decide, the use of a LR may also be the legally correct option [5 – 6].

Apart from its correctness, the LR approach has several other important properties. It allows, for example, the combination of evidence of different types, nicely demonstrated in the testing of both automatic and acoustic-phonetic features in hybrid FVC systems. Requiring as it does an estimate of the probability of the evidence under both defence and prosecution hypotheses, it also allows an expert in an adversarial system to be impartial. Finally – a crucial desideratum in forensic comparison science – the accuracy (and more recently the precision) of a LR-based FVC system are also straightforwardly tested [7].

As far as practice is concerned, a LR approach has also been implemented as a matter of course since the mid 1990's in DNA profiling [8] (although it is not clear that the courts actually understand that the random occurrence probability usually quoted is part of a LR). The use of LRs in forensic voice comparison was an idea whose time came around the beginning of the new millennium, when its efficacy first began to be demonstrated both with automatic and traditional, usually acoustic-phonetic, features [9]. The results from now well over a decade's extensive, and continuing, research testing with both automatic and acoustic-phonetic features have shown that the approach works rather well, in the sense that same-speaker speech samples can be rather well discriminated from different-speaker speech samples on the basis of the magnitude of their LRs e.g. [10], and it has been shown in [11] that the approach can emulate the DNA gold-standard, a desirable fact, given that "… DNA profile evidence is now seen as setting a standard for rigorous quantification of evidential weight that forensic scientists using other evidence types should seek to emulate" [3]. The LR-based testing of other forensic evidence types is now following: fingerprints [12], handwriting [13] and SMS texting [14].

Despite these encouraging developments in FVC, which have been interpreted as an incipient paradigm shift towards a maximally objective evaluation of forensic evidence [9], it is safe to say the idea has not yet caught on. There are probably many reasons for this. The legal profession is, firstly, inherently conservative and thus adheres to the (surely intuitive) assumption that the forensic expert should be asked to estimate the probability of a hypothesis, given the evidence. For example, in a recent draft proposal of standards for the interpretation of forensic evidence, representatives of organizations including the Australian Attorney-General's Department, the Australian and New Zealand Forensic Science Society, the University of New South Wales' Expertise, Evidence and Law Program, and the New South Wales Bar Association implicitly upheld the first type of FVC in recently maintaining that "Interpretation [of evidence] includes answering the question as to whether or not …items share a common origin…" [15]. Secondly, it appears, many legal practitioners find it difficult to understand the LR approach if they actually encounter it. This can be seen most

clearly in inconsistencies in statements concerning likelihood ratio evaluation. For example, the aforementioned Australian draft standards proposed a probability-of-hypothesis-given-evidence approach while simultaneously recommending a text-book on the interpretation of evidence [5] which clearly espouses the opposite, probability-of-evidence-under-competing-hypotheses likelihood-ratio approach. The degree of misunderstanding by the courts is probably best exemplified by the mind-boggling confusion and inconsistencies, described in detail in [16], in a 2010 ruling by the England and Wales Court of Criminal Appeal in *R v T* [17].

It is not just the legal profession that finds the approach difficult to understand, however. A recent book on Forensic Linguistic evidence misrepresents LRs [18], and perhaps the best-known phonetics text-book [19] defines the LR as "… the likelihood that the two voices in question are the same as compared with the likelihood that they are different." thus confusing it with the prior odds (the same mistake is found in the *R v T* ruling [16]). Steam-engine time[1] for the likelihood ratio approach in FVC still looks a very long way off.

I started to use LR estimates in my case-work in 2002 (some details from actual cases are in [21 – 22]), but the case described here is to my knowledge the first in Australia where LR-quantified speech evidence was actually received in court (there is to date only one other). Australia's legal system is adversarial and I appeared for the prosecution. Under the assumption that a "logically incorrect conclusion that's 'understood' is no alternative to a logically correct conclusion which needs explanation" [23], one aim of this paper is therefore to briefly document the first FVC case where LR-based speech evidence was received in court in Australia, and to try thereby to give an idea of what is involved. I will also briefly address areas where the LR estimation might have been substantially improved, given what we know from research in the interim. A third problem is also broached, concerning the problems of conveying the meaning of LRs to a jury, and integrating any LR-based evidence with the other evidence in a case.

2 How to make $150 million in one phone call

On Christmas Eve 2003 a fraudulent fax was sent to the investment bank JP Morgan Chase in Australia requesting the transfer of $150 million to accounts in Switzerland, Greece and Hong Kong. About 10 minutes before the closing of business, the bank received a phone call from a Craig Slater, asking for a call-back on the fax (a procedure confirming the details of the fax and verifying that the transfer could go ahead). The phone call started with the following exchanges (E = J.P. Morgan employee, F = fraudster):

E *J.P. Morgan Greg speaking*
F **yeah hello Greg this is Craig Slater here mate**

[1] 'Steam engine time' is a metaphor for the widespread acceptance of an idea or invention which, like the steam engine, has made a sudden polygenetic emergence after actually being around for a long time [20].

E *oh g'day how are you*
F **not too bad I've been having a bit of trouble here we erh I haven't been able to get onto anyone else on the other lines …**
E *yep*
F **is it would it be possible to you to do a call back**
E *erh just a second I'll just go check the fax*
F **O. … O.K.**

The bank employee then reads out from the fax the amounts to be transferred, and Slater tersely confirms them, for example:

E *erh and we're going to pay Hong Kong dollars one one eight six seven eight five four three spot two nine* [$118,678,543.29] *to HSBC erh Hong Kong*
F **correct**
…
E *Hong Kong I think Hong Kong Power Limited six three six double oh three oh five five double oh one* [$636,003,055,001]
F **yes**

The call then ends with the appropriate season's greetings:

E *is that correct*
F **that's correct**
E *OK then*
F **than .. thank you very much**
E *have a good Christmas*
F **you have a good Christmas too. bye**
E *OK bye*

The Australian Commonwealth Superannuation Scheme account administered by the bank was now $150 million short.

3 Approach

When using traditional features to estimate a LR, one first scours the offender and suspect data for comparable material, e.g. same utterance; same phonemes in comparable environments. Although the fraudulent call lasted just over 3.5 minutes, it contained only about 14 seconds of offender speech, and much of that lacked material useful for acoustic-phonetic voice comparison. There were, however, five repetitions of the word *yes*, and these could be compared with many tokens of the word *yes* in several recordings of the suspect during previous police and customs interviews. (It is actually unusual to encounter *yes* in forensic speech material: people usually say *yeah*.) In addition, the fraudster's utterance *not too bad* was to prove useful in the light of its occurrence, with the same intonation, in recordings of telephone intercepts of the suspect talking to his mates. The bulk of the LR estimate in this case thus rested on features extracted from *yes* and *not too bad*.

In order to estimate the probability of getting the difference between the offender sample features and another person chosen at random from the relevant population (the LR denominator) a reference sample is of course necessary. In this case, no explicit alternative hypothesis having been nominated by defence, it was sensible to assume that the offender voice belonged to a male speaker of Australian English, and I amassed a sample of 35 adult males with ages uniformly distributed between early 20's and early 70's who agreed to be phoned up and recorded. Unlike, for example, *I'll blow yer fuckin' head off*, the test material in this case has the enormous benefit of easy ecologically valid elicitation: reference sample speakers were instructed to appropriately reply *not too bad* and *yes* or *no* to a set of questions I asked, like *how's it going*? or *are you inside*? I tried to indirectly prime reference subjects once at the beginning of the elicitation simply by saying *not too bad* with the correct intonation, but they were not explicitly told the correct intonation to produce. In this way it was possible to obtain several replicates of *yes*, and *not too bad* with the correct intonation, from most speakers in the reference sample. To sample non-contemporaneous within-speaker variation, subjects were phoned and recorded on several occasions over the course of several weeks.

4 F0 in *not too bad*

Although parameters of long term F0 distributions have been shown to perform rather well in LR-based discrimination research [24], F0 is usually not of much use in real case-work because of its disadvantageous short-term variance ratio. This case was different, however, in that all test data *not too bad*s are said in response to the interlocutor's asking how the suspect/offender is. They have very similar intonational structure typical for this conversational interchange, conveying the typical rise nuclear tone meanings of supportive interest encouraging further conversation [25]. They have a rising nuclear tone on *bad*, realised with a low rising allotone. The offender *not* carries a high head realised with a high, slightly falling pitch[1] (probably induced from the coda stop); the suspect's *not* has either a high head realised in the same way, or a low head, with a low pitch. The pitch on *too* is interpolated in the expected way. This near identical intonational structure means that the F0 values on *not too bad* are highly comparable and might be expected to yield useful likelihood ratios (i.e. values that deviate substantially from unity). Figure 1 shows the F0 realising the [H.L.LH] intonational pitch of the offender, aligned with its wideband spectrogram. F0 on *not* can be seen to drop from about 200 Hz to 175 Hz; whence it drops further on the nucleus of *too* to about 125 Hz. The F0 shows a small ca. 15 Hz increase from its minimum value of 125 Hz in the /b/ hold, and rises on the nucleus of *bad* with a slightly convex contour from about 145 Hz to peak at about 185 Hz. Figure 2 compares the offender F0 with the F0 of the suspect's 15 *not too bad* utterances. The similarity is considerable, with the offender's F0 time-course lying completely

[1] The term *pitch* is commonly found referring to both the perceptual property of the linguistic category of intonation and its acoustical reflex of F0. In this paper it has the former reference only.

within, and in some places almost exactly in the middle of, the suspect's distribution. Note too the suspect's use of both H and L on *not*.

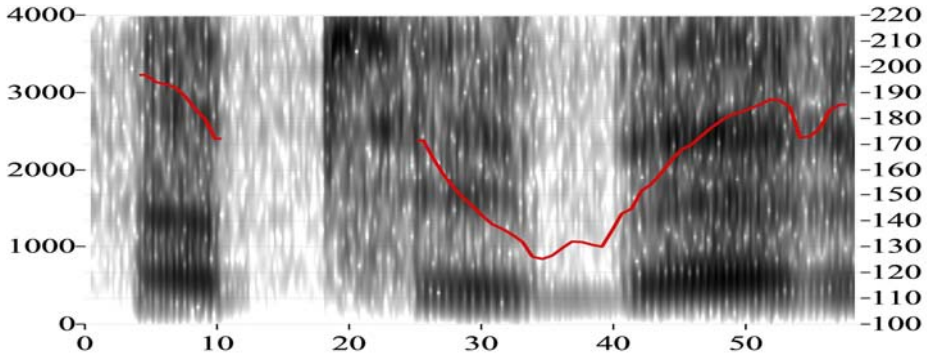


Fig. 1. F0 time course of offender's *not too bad* (right-hand scale in Hz) superimposed on wide-band spectrogram (left-hand scale in Hz). x axis = duration (csec.).

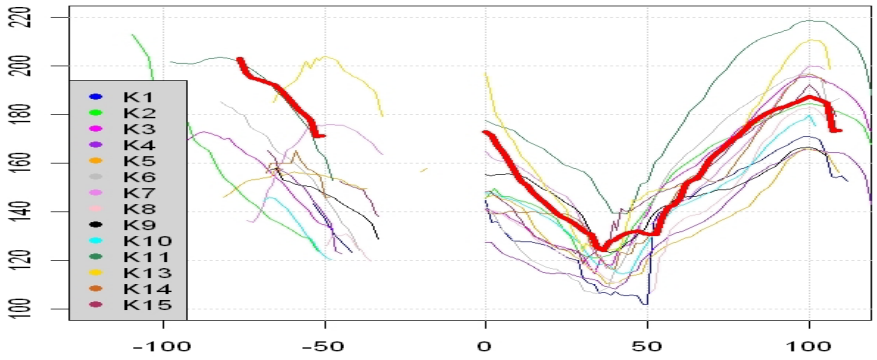


Fig. 2. Suspect (K) and offender (red) *not too bad* F0 plotted as function of equalized duration. Axes: vertical = F0 (Hz), horizontal = duration equalized around /u:/ onset (0%) and peak F0 in bad (100%).

The *not too bad* F0 values from the reference sample speakers showed no surprises in variance structure. Figure 3, which plots the F0 time-course for just two of the reference speakers on two different occasions, illustrates this. There was, firstly, expected within-speaker variation in phonological structure between a H and a L tone on *not*, just as in the suspect. The first speaker in figure 3 displays this behaviour: in recording sessions 1 & 2 the different F0 height on his *not* corresponding to the H vs. L distinction can be easily seen. This speaker also shows considerable non-phonological variation in F0 between sessions: his *not too bad*s in session 2 are much higher and sound more enthusiastic that in session 1, for example. A comparison with the second speaker in figure 3 illustrates typical aspects of between-speaker variation.

Most obviously, he shows different overall F0 values from the first speaker. But the two speakers also differ in their within-speaker variation: unlike the first speaker, the second speaker always had a H tone on *not* and shows little difference from session to session.

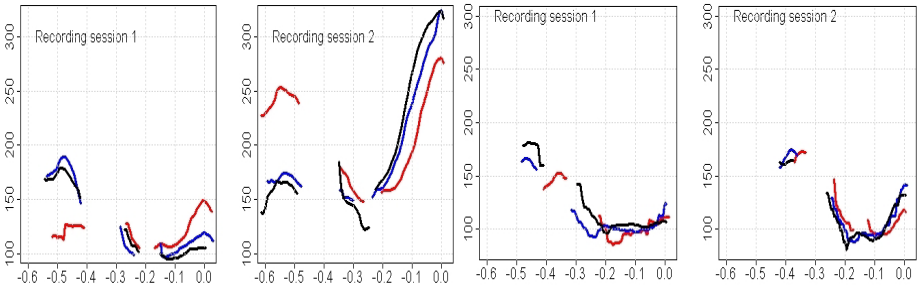


Fig. 3. Within- and between-speaker variation in reference sample: *not too bad* F0 tokens from two speakers on two different occasions. Axes: vertical = F0 (Hz), horizontal = duration (csec. from peak F0 in bad).

The F0 time course in *too bad* was sampled at four points: mid point of the vowel in too, and at first target, peak and mid way between in the vowel in *bad*. Because of the aforementioned associated phonological variation between H and L, F0 on not was not used. The four sampled F0 values were treated as multivariate data and LRs were estimated using a generative multivariate LR formula [26], modeling the reference sample both normally and with a kernel density. The difference between the suspect and offender F0 values in *too bad* was estimated at about 20 times more likely assuming they had come from the same speaker, irrespective of the normal or kernel modeling of the reference sample. Given the high degree of similarity between suspect and offender samples (figure 2), the relatively low LR value for the comparison is salutary and reminds us that evidentiary value is also dependent on typicality.

5 Formant pattern in *yes*

Both suspect and offender *yes* tokens sounded to have a more centralized /e/ than normal, and lower than normal pitch on /s/ (somewhat reminiscent of the pharyngealised /s/ in Arabic). Figure 4 shows spectrograms with superimposed formants of two *yes* tokens from offender and suspect. Both show unremarkable F-pattern configurations given the auditory impression – note the lower than normal cut-off frequencies for the /s/ (four formants were extracted below 3.6 kHz, so the F-pattern of the /s/ is not tracked well). The offender and suspect /je/ F-pattern was sampled at onset, midpoint and offset. Figure 5 shows that the sampled F-pattern values of the suspect and offender are fairly similar, the offender's values lying almost totally within the distribution of the suspect. Thus the probability of getting the offender values assuming they have come from the suspect will be fairly high. Figure 5 also shows the distribution of the reference sample values for /je/. It can be seen that the between-speaker

variance in F2 and F3 is quite big, but more importantly it shows that the F-pattern in the suspect and offender has unusually low offset values. This presumably correlates with the audibly lower, more centralized nucleus. This means that a LR greater than unity is to be expected. A MVLR was estimated for the F2 and F3 vales at all three sampling points in /je/, as it was assumed that the F1 would have been compromised by the telephone transmission. Both normal and kernel density MVLRs indicated that the difference between suspect and offender /je/ F-pattern was about 70 times more likely had they come from the same rather than different speakers.

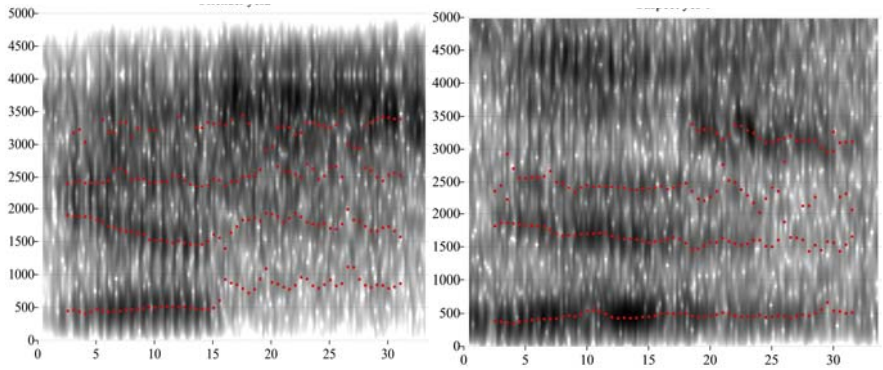


Fig. 4. Spectrograms of offender (left) and suspect (right) *yes* tokens. x axis = duration (csec.)

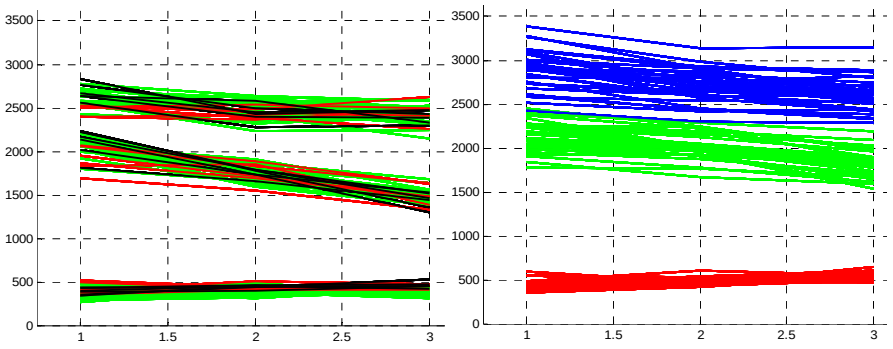


Fig. 5. /je/ F-pattern values in *yes* sampled at onset, midpoint and offset. Left: individual tokens from offender (black) compared with suspect (red, green). Right: mean values from reference sample. Axes: vertical = frequency (Hz), horizontal = equalized duration.

6 Other features

All segmental aspects of the *not too bad* acoustics are also theoretically comparable, and LRs for the F-pattern on its three constituent vowels (/o/ in *not*, /u:/ in *too* and /æ/ in *bad*) were also estimated. As expected, comparison with /o/ was complicated due

to nasal poles associated with quasi-Helmholtz and nasal cavity resonances induced by /n/. Figure 6 compares the F-pattern of offender and suspect in these three vowels (nasal poles in /o/ are not shown). It can be seen that for all three vowels there is, as with the F0, considerable similarity, the offender values lying almost totally within the distribution of the suspect's values. The F-pattern differences were parametrised by mean values taken over quasi-steady-state portions at vocalic target, and point measurements at offset. Figure 7 shows some of these values against the corresponding reference sample distributions, where it can be seen that they are fairly typical (as in *yes*, some values were discarded because of possible influence from the telephone transmission). The resulting MVLR estimates were all greater than unity: ca. 24, 5 & 11 for /o u: æ/ respectively. In addition, I attempted a crude (because discrete) LR estimate for the low cut-off in the /s/ spectrum in *yes*: this showed that a low cut-off was conservatively about 2.5 times more likely in both samples if they had come from the same rather than different speakers.

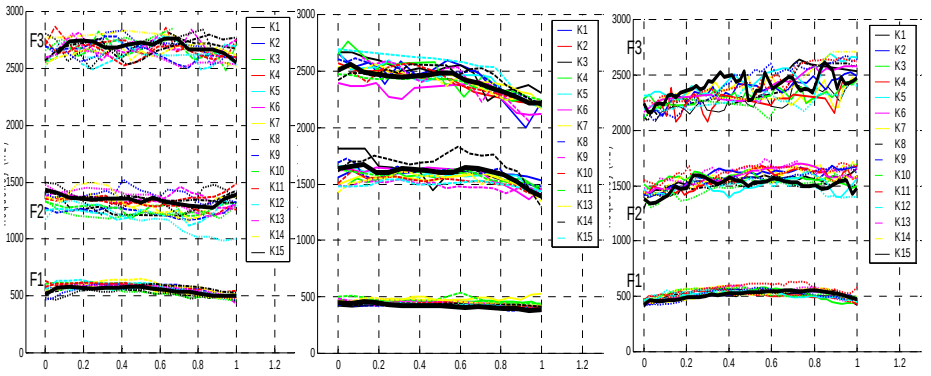


Fig. 6. Offender's F-pattern (thick line) plotted against suspect's values in the vowels of *not* (left) *too* (middle) and *bad*. Vertical axis = frequency (Hz), horizontal axis = equalized duration (% * .01).

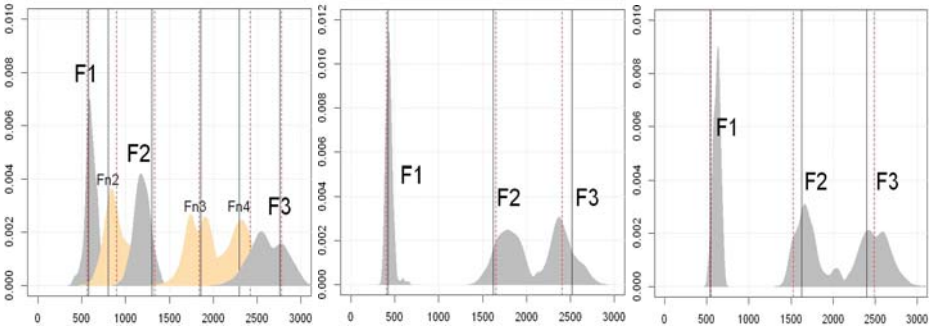


Fig. 7. Target F-pattern means for offender (vertical dotted red lines) and suspect (vertical solid black lines) in the vowels in *not* (left), *too* (middle) and *bad* plotted against the reference sample distributions. Fn = nasal pole. Vertical axis = probability density, horizontal axis = frequency (Hz).

7 Result

A naïve Bayes combination of the LRs from the features described yielded an exceptionally large overall LR of about 11 million (one feature not described contributed an additional LR of ca. 2). Since I suspected some between-segment correlation, but could not estimate it, I simply discarded the putatively correlated LRs (e.g. from individual formants in *not too bad*) to arrive at a much smaller LR estimate of ca. 300,000. According to Bayes' Theorem there would now have to be more than roughly 16,000 others who could have said the incriminating speech before the posterior probability that the suspect said it fell below 95%.

8 Critique

The now five years since this case have seen continuing improvements in traditional FVC LR estimation. The main ones have been in the recruitment of various aspects of automatic FSR approaches to (1) better parametrise traditional features like formants [27]; (2) combine LRs from different features with logistic-regression-fusion, thus providing a potential solution to the problem of possible between-segment correlation [28]; (3) use features like cepstral coefficients as well as formants to characterize segments [29], and (4) provide measures of accuracy and precision [7]. All of these would probably have helped enormously in this case. Firstly, given the segmental and suprasegmental identities in *not too bad* and *yes*, it would have obviously been advantageous to parametrise both F-pattern and F0 trajectories with DCT or polynomial coefficients, rather than use point estimates. Secondly, the problems with handling possible between-segment correlation might have been obviated by logistic-regression fusion of the LRs from the different segments. Finally, the whole of the /s/ spectrum in *yes* could have been compared using cepstrally-mean-subtracted CCs, rather than just the fact that it had a low cut-off. The effect of incorporating these methods is an interesting and empirical one, and I am currently re-estimating the LR for the evidence with them to see what sort of a difference they make. At the time of the trial, however, to which we now turn, these improvements lay in the future.

9 Reception

The final point in the FVC process was of course the trial. This where the science ends and the law begins. My perception of how the LR-based evidence was received in court is as follows (one juror, who received judicial admonishment for continuously falling asleep, obviously didn't receive much). Prosecution and defence strategies to lead/attack my evidence had to be based on their assessment of what was best going to convince a jury, and this is almost certainly not going to be academic arguments relating to the strengths and weaknesses of the LR approach. Prosecution

seemed to put emphasis on my showing a redundantly large number of spectrograms, perhaps thereby trying to emphasise the idea to the jury that the approach was scientific. Defence suggested the LR approach was somehow my personal development, and by implication therefore not widely accepted. They also argued that in selecting the reference sample I had not taken into account the fact that the suspect was from a particular area of Sydney. This is a spurious argument. In asking what the probability is of getting the offender data assuming it had come from the suspect, one does indeed partially condition on the suspect. However, the reference sample is chosen with respect to the defence hypothesis, and that must sensibly relate to the offender's voice, not the suspect's (and in any case the features used were not such as to be expected to vary across different areas of Sydney).

I felt that it was not wise to try to explain such things to a jury. Instead I tried to concentrate on emphasizing two points. Firstly, that I was trying to estimate the strength of the evidence and not the probability that the suspect said the incriminating speech. Secondly, that the jury should not give much weight to the precise value of the LR; only that it was very big. It is encouraging to report that I felt tremendously aided in this by the judge, who insisted that I repeat these ideas many times, so that the jury might have a chance of understanding them.

A well-known probability expert has cautioned: "… one must be wary of oversimplistic direct interpretation of the numerical value of the likelihood ratio, which can only be sensibly considered in conjunction with other information"[30]. However, no attempt was made in the trial to explain to the jury the meaning of the LR-based voice evidence by introducing this extra information – the notion of combining it with prior odds to estimate a posterior. From a theoretical point of view, of course, this is a major problem with the use of a LR-based approach, since the magnitude of the evidential LR has no meaning outside of its matrix theorem (e.g. with prior odds of 1 to 100 against, a LR of 1000 means a posterior probability of ca. 91% in favour; but with priors of 1 to 1000, the same LR gives a posterior of only 50%). This means that proposals to explain LRs verbally in terms of varying degrees of strength of evidence *in support of the defence or prosecution hypotheses* are misleading: a LR *indicating a high degree of support for the prosecution* can be transformed into a posterior supporting the defence by sufficiently large priors, and *vice-versa*. In the reality of a trial, however, there remain big problems with trying to make sense of strength of evidence estimates in terms of Bayes' Theorem (quite apart from the fact that the way it works is not well understood by the advocates whose job it would be to explain it). The main problem was pointed out by Justice Hodgson [31] – a rare combination of appeals court judge and probability expert, and thus worth heeding. It is simply that since not all types of evidence in a trial can be sensibly assigned a LR there is no way of combining à la Bayes the LR-based evidence with the non-numerically based evidence. The law expects the jury, after all, to evaluate the evidence using their commonsense. As Hodgson also points out, this is one reason why another popular attempt to explain the meaning of the LR to the jury – e.g. *whatever your belief in the hypothesis was before the evidence you must increase/decrease it by the amount of the LR* – is also not going to work.

We thus end up, at least in LR-based FVC, with a current impasse at the boundary between science and the law. But it is an impasse, apparently, that is unsatisfactory and frustrating only from the point of view of the scientist. I have no idea, of course, of what sense the jury made of my LR-based voice evidence. Whichever way it was construed, and combined with the other evidence, they returned a verdict of guilty [32 – 33].

10 References

1. http://news.bbc.co.uk/2/hi/uk_news/england/west_yorkshire/4039057.stm
2. Aitken, C.G.G., Taroni, F.: Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, Chichester (2004)
3. Balding, D.J.: Weight of Evidence for Forensic DNA Profiles. Wiley, Chichester (2005)
4. Association of Forensic Science Providers: Standards for the formulation of evaluative forensic science expert opinion. Science & Justice 49, 161 – 164 (2009)
5. Robertson, B, Vignaux, G.A.: Interpreting Evidence – Evaluating Forensic Science in the Courtroom. Wiley, Chichester (1995)
6. Morrison, G.S.: Forensic Voice Comparison. In: Freckleton, I., Selby H. (eds.) Expert Evidence. Ch.99 (2010)
7. Morrison, G.S.: Measuring the Validity and Reliability of forensic likelihood-ratio systems. Science & Justice (51) 3, 91-98 (2011)
8. Forman, L.A., Champod, C., Evett, I.W., Lambert, J.A., Pope, S.: Interpreting DNA evidence: a review. International Statistics Journal 71 473 – 495 (2003)
9. Morrison: G.S. Forensic voice comparison and the paradigm shift. Science & Justice 49, 298–308 (2009)
10. Gonzalez-Rodriguez J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech and Language 20, 331 – 355 (2006)
11. Gonzalez-Rodriguez J., Rose P., Ramos, D., Torre, D. & Ortega-García, J.: Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. IEEE Trans. on Audio Speech and Language Processing 15(7), 2104 – 2115 (2007)
12. Neumann, C., Evett, I.W., Skerrett, J. Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. Journal of the Royal Statistical Society 175, 371 – 415 (2012)
13. Hepler, A.B., Saunders, C.P., Davis, L.J., Buscaglia, J.: Score-based likelihood ratios for handwriting evidence. Forensic Science International, (2012)
14. Ishihara, S.: A Forensic Text Comparison in SMS Messages: A Likelihood Ratio Approach with Lexical Features. In: Clarke, N., Tryfonas, T., Dodge, R. (eds.) Proc. Seventh Intl. Workshop on Digital Forensics and Incident Analysis, 55 – 65 (2012)
15. Standards Australia: Forensic Analysis Part 3: Interpretation. Draft for Public Comment, DR AS 5388.3 http://infostore.saiglobal.com/store/Details.aspx?ProductID=1530224 (2012)
16. Morrison, G.S.: The Likelihood Ratio framework and forensic evidence in court: A response to *R v T*. International Journal of Evidence and Proof 16, 1 – 29 (2012)
17. EWCA 2439 (2010)
18. Morrison, G.S.: Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. Australian Journal of Forensic Sciences 41, 155 – 161 (2009)

19. Ladefoged, P: A Course in Phonetics. 5th ed. Thomson, Boston (2006)
20. Pratchett, T., Stewart, I, Cohen, J.: Darwin's Watch. Ebury Press, London (2005)
21. Rose, P.: The Technical Comparison of Forensic Voice Samples. In: Freckleton, I., Selby H. (eds.) Expert Evidence. Ch.99 (2002)
22. Rose, P.: Technical forensic speaker recognition: Evaluation, types and testing of evidence. Computer Speech and Language 20, 159 – 191 (2006)
23. Berger, C.: Criminalistics is reasoning backwards. Nederlands Juristenblad AFL13 (2010)
24. Kinoshita, Y., Ishihara, S., Rose P.: Exploring the Discriminatory Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition. Intl. J. Speech Language & the Law 16(1), 91 – 111 (2008)
25. Wells, J. C.: English Intonation. Cambridge University Press, Cambridge UK (2006)
26. Aitken, C.G.G., Lucy, D.: Evaluation of trace evidence in the form of multivariate data. Applied Statistics 53(4), 109 – 122 (2004)
27. Morrison, G.S.: Likelihood Ratio forensic voice comparison using parametric representation of the formant trajectories of diphthongs. JASA 125, 2387 – 2397 (2009)
28. Pigeon, S., Druyts, P., Verlinde.: Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. Digital Signal Processing (10) 1-3, 237 – 248 (2000)
29. Rose, P.: Forensic Voice Comparison with Secular Shibboleths – a hybrid fused GMM-Multivariate Likelihood Ratio-based Approach Using Alveolo-Palatal Fricative Cepstral Spectra. Proc. ICASSP, 2011.
30. David, P.: Statistics and the Law. In: Bell, A., Swenson-Wright, J., Tybjerg, K. (eds.) Evidence. Cambridge University Press, Cambridge UK, 119 – 148 (2002)
31. Hodgson, D.: A lawyer looks at Bayes' theorem. The Australian Law Journal 76, 109 – 118 (2002)
32. http://www.abc.net.au/pm/content/2008/s2451596.htm
33. Commonwealth Director of Public Prosecutions Report 2009-2010.

Keynote talk

Title: Google's speech internationalization project: From 1 to 300 languages and beyond

Abstract:

The speech team at Google has built speech recognition systems in more that 30 languages in little more than 2 years. In this talk we will describe the history of this project and what technologies have been developed to achieve this goal. I'll explore a bit some of the acoustic modeling, lexicon, language modeling, infrastructure and even social engineering techniques used to achieve our ultimate goal, to build speech recognition systems in the top 300 languages of the planet as fast as possible.

Speaker Bio:

Dr. Pedro J. Moreno leads the speech internationalization engineering group at the Android division of Google. His team is in charge of the infrastructure, engineering, and research needed to deploy and maintain multilingual speech recognition services worldwide.

He joined google 8 years ago after working as a research scientist at HP Labs. During his work at HP he worked mostly in audio indexing systems.

Dr. Moreno completed his Ph.D. studies at Carnegie Mellon University under the direction of Prof. Richard Stern. His work there was focused on noise robustness in speech recognition systems.

His Ph.D. studies were sponsored by a Fulbright scholarship.

Before that he completed an Electrical Engineering degree at Universidad Politecnica de Madrid, Spain.

Poster Session P1

Audio Encoding for Heart and Breath Sounds Acquired with Digital Stethoscope

Ignacio Palacios Santos[1], Doroteo Torre Toledano[1] and Andrés Martínez Fernández[2]

[1] ATVS, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, SPAIN
ignacio.palacios@estudiante.uam.es, doroteo.torre@uam.es
[2] Universidad Rey Juan Carlos, Fundación Enlace Hispano Americano de Salud, Fuenlabrada, SPAIN
andres.martinez@urjc.es

Abstract. This paper presents a real-time audio encoding system for cardiorespiratory sounds acquired with a low cost Bluetooth digital stethoscope developed by EHAS (Enlace Hispano Americano de Salud) Foundation and Fundatel Foundation. The system is currently working with a G.722 speech coder for compression and subsequent transmission of cardiorespiratory signals. However, these signals have different frequency characteristics from speech. Therefore, we sought a better adapted alternative to encode these signals optimally, with an encoder not subject to the payment of any license, with low computational cost, good quality and low bandwidth. We have evaluated the proposed solution both using objective measures such as root mean squared error and using subjective opinions from four expert physicians.

Keywords. Cardiorespiratory Sounds, Stethoscopy, ADPCM Encoding, Development Cooperation, VoIP.

1 Introduction

EHAS Foundation is a non-profit organization which aims to promote the appropriate use of new Information and Communication Technology (ICT) to improve health processes in isolated rural areas of developing countries. In 2011, a pilot project was conducted in the Peruvian area of River Napo, with the objective of providing cardiorespiratory tele-diagnosis for the local population. The diagnosis is made between rural health posts (without a physician) and health centres of reference (with medical presence). With the audio encoding system, the physician is able to guide the patient via videoconference by means of a digital stethoscope at each end, transmitting the cardiorespiratory audio over VoIP. This solution will allow many isolated people to have a quick diagnosis, sparing them the travelling of hundreds of miles to get to the nearest hospital, going through poorly served areas. EHAS Foundation, together with Fundatel Foundation, developed a low cost bluetooth digital stethoscope and a peer-to-peer transmission system for the cardiorespiratory signals. This paper sets forth a cardiorespiratory real-time audio encoding alternative for digitalized signals that is

able to send these signals over VoIP with good quality, in real-time and using much less bandwidth than previous solutions available at EHAS Foundation. The development of a digital tele-stethoscope is not a new idea in itself [1, 2]. This work does not focus in the development of the digital tele-stethoscope, but in the optimization of the audio coding for the cardiorespiratory signals involved in digital tele-stethoscopy.

2 Characterization of Cardiorespiratory Signals

Cardiorespiratory sounds are obtained through auscultation which is the term used for the listening of internal sounds of the body, usually by means of a stethoscope. Auscultation is performed for the purpose of examining the circulatory system and respiratory system (heart and breath sounds). There are four auscultation areas: mitral, aortic, tricuspid and pulmonary. The sounds detected in these four areas will have different frequency characteristics than voice signals.

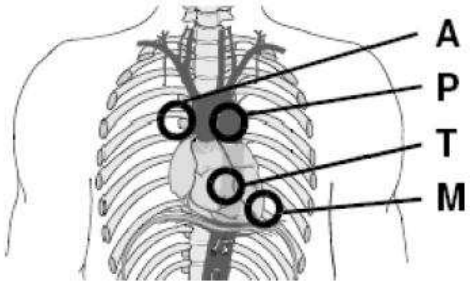


Fig. 1. Auscultation areas [3]

Generally, the components of heartbeat and lung sounds that are useful for diagnostic purposes are in the range of 20-1000 Hz. The first heart sounds (S1 and S2), are produced by the closing of the atrioventricular valves and semilunar valves respectively. These sounds fall in the range of 20-115 Hz. Disorders such as heart murmurs occur in the range of 140-600 Hz [4]. Thus, a suitable listening range for heart sounds would be approximately 20-600 Hz. For respiratory sounds, the strongest part of the signal is typically under 100 Hz, although the signal can have useful components up to 1.2 kHz [5].

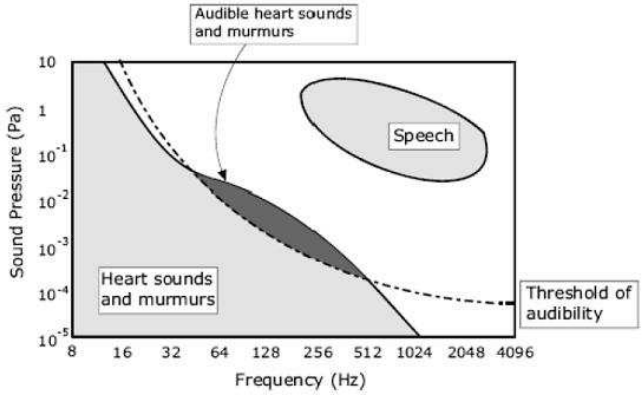


Fig. 2. Cardiorespiratory sound frequency distribution [3]

Hence, by the Nyquist theorem [6]:

$$f_s > 2f_M \tag{1}$$

we could derive the minimum sampling frequency (f_s) to which we can downsample cardiorespiratory signals without losing relevant information for the diagnosis. In our case the maximum frequency present in the signal (f_M) is 1200 Hz. Hence by Nyquist theorem the cardiorespiratory signas can be completely recovered from samples at a sampling frequency (f_s) higher that 2400 Hz.

3 EHAS-Fundatel System Architecture

As mentioned in [3], the analogical cardiorespiratory sound captured by the chestpiece of the digital Bluetooth stethoscope EHAS-Fundatel is picked up by a microphone, which converts the mechanical motion of the sound pressure waves into a voltage variation. After the analog signal is captured, the following process is applied:

1. This voltage is transmitted, after an analogical amplification stage, to a TLV320AIC33 hardware codec. The codec hardware, at its first stage, performs the conversion of the analog signal emitted by the microphone to a digital signal that can be processed by other electronic devices. This signal is downsampled to 8 kHz and 16 bits per sample.

2. The resulting signal is sent to a OEMSPA310 Bluetooth chip that sends the audio to a PC using the Serial Port Profile (SPP) Bluetooth protocol. This audio is sent encoded in a format that adapts the sound to the features of Bluetooth transmission. During this communication, flow control information is also sent, allowing the PC to ensure it is receiving the audio without any alteration caused by a poor

management of the speed of transmission/reception, as well as the necessary information to ensure the audio can be rebuilt.

3. There is a Stethoscope's software [3] that performs a pre-encoding protocol before sending it to a VoIP application (Ekiga Softphone [7]) when the digital sound reaches the PC. The aim is to support any sound server, running at any sample rate and adjust the sending rate of data at 44100 samples per second and 32 bits per sample (which is the most common case).

4. Once the signal reaches the softphone, the latter adapts the digital audio input to the needs of the audio codec used (G.722 in our example, which requires an entry with 16000 samples per second and 14 bits per sample, providing an output of 64 Kbps).

5. After sending the final digitalized audio to the network, the audio is received by the remote computer, starting the reverse process (decoding).

4 Encoders Test

In order to find an alternative to G.722 codec several simulations of EHAS-Fundatel system architecture were performed with different encoders. We used 33 audio cardiorespiratory samples, 15 of which were obtained with the digital stethoscope's bluetooth EHAS-Fundatel (3 mitral-type, 3 aortic-type, 3 tricuspid-type and 6 pulmonary-type), and 18 downloaded from [8] (10 cardiac-type, 8 pulmonary-type). We created several Linux scripts where the audio samples were encoded and decoded using the audio manipulation software SoX [9] (Sound eXchange), obtaining different figures of merit as results that allow objective comparison of the encoders. These figures of merit are Compression Factor (CF), Bit Rate (R), Root Mean Square Error (RMSE), Percent RMS Difference (PRD) and Real Time Factor (RTF):

$$CF = \frac{source_data_size}{encoded_data_size} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(x[n]-\tilde{x}[n]\right)^2}{N}} \tag{3}$$

$$PRD = \sqrt{\frac{\sum_{n=1}^{N}\left(x[n]-\tilde{x}[n]\right)^2}{\sum_{n=1}^{N}x^2[n]}} \times 100\% \tag{4}$$

$$RTF = \frac{computation_time}{audio_duration} \quad (5)$$

Where $x[n]$ is the original signal and $\tilde{x}[n]$ is the signal obtained after coding and decoding the audio, and N is the number of samples in the audio.

The EHAS-Fundatel audio samples are digitalized at 44.1 kHz and 32 bits per sample, while the heart sounds [8] are digitalized at 16 kHz and 16 bits per sample; finally respiratory signals are digitalized at 11.025 kHz and 16 bits per sample. Initially we tested with G.722, G.726, IMA-ADPCM encoders and downsampled at 1, 2 and 4 kHz.

Table 1. Comparison encoders assuming a 16 bit per sample at 16 kHz input.

	CF	R (Kbps)	RMSE	PRD (%)	RTF
G.722	4	64	0.039405267	22.7856303	0.2122
G.726-2	16	16	0.004458867	3.22418553	0.03503
G.726-3	10.66	24	0.002614943	1.93012648	0.03636
G.726-4	8	32	0.002327377	1.71252903	0.03420
G.726-5	6.4	40	0.003868983	2.84869895	0.03633
4 kHz	4	64	0.000347993	0.29957613	0.00448
2 kHz	8	32	0.002195643	1.98224353	0.00446
1 kHz	16	16	0.010389677	5.048618	0.00441
IMA-ADPCM	4	64	0.000396343	1.93784919	0.00613

The G.726 codec worked well, but it requires as input audio previously encoded with G.711 μ-law or A-law. Furthermore, the decoded audio seemed to have worse subjective quality. The G.711 codec is a simple PCM with minimum processing load on the CPU of the PC, and its use is not subject to the payment of any license. However, the quality of the received audio is very dependent on the level of pressure applied to the chestpiece. This codec uses an 8-bit not uniform quantification, so that low amplitude signals suffer less distortion than those with large amplitude. As the cardiorespiratory signal presents large amplitude on low frequency components, this requires auscultation under low pressure levels, preventing signal distortion caused by the encoder working in its less linear region. In addition, working under low chestpiece pressure levels means, in practice, using a smaller number of bits per sample encoded, so the resolution is severely affected.

IMA-ADCPM codec and downsamplings had better results than the G.722 regarding the RMSE, PRD and RTF figures of merit, so we decided to design an integrated

codec consisting of an IMA-ADPCM codec previously downsamped at 4 kHz, thus respecting the bandwidth of cardiorespiratory signals.

Table 2. Comparison between the current codec (G.722) and alternative (Proposed Stethoscope-Codec)

	CF	R (Kbps)	RMSE	PRD (%)	RTF
G.722	4	64	0.039405267	22.7856303	0.02122
Proposed Stethoscope-Codec	16	16	0.002175277	1.93784919	0.00588

The new Proposed Stethoscope-Codec provides better objective results than the G.722 codec when comparing all figures of merit. In particular, provides better objective quality with four times less bit rate and less computational complexity.

5 Medical Evaluation

After performing the objective assessment of the encoders, a subjective evaluation of these was carried out, since it is intended that the encoding of the cardiorespiratory audio does not distort too much the quality of the audio signal and that the decoded audio is useful for remote diagnosis. The 33 original samples and the respective samples decoded with the G.722 codec and Stethoscope-Codec were evaluated by a group of four physicians of the Hospital Clínico San Carlos in Madrid. The doctors were asked to grade from 1 to 10 the quality of the audio heard, both the original and decoded without knowing each sample's encoding. These physicians were three pulmonologist and one internal medicine doctor.

Only one doctor graded, in general, the new Proposed Stethoscope-Codec with a higher grade, stating that there was less noise in the audio heard and that, in such a way, the cardiac and respiratory audio where she needed to focus were highlighted. The rest of the physicians gave, in general, a slightly better grade to the G.722 codec with respect to the new Proposed Stethoscope-Codec. Other doctor explained that in the case of samples related to pathologies where very high-frequency noises were detected, a distortion was observed with such noises taking a deeper register. Nevertheless, being familiar with the pathology, it was possible to diagnose it.

Table 3. Average grades for 33 samples (1 to 15 EHAS-Fundatel samples, 16 to 25 cardiac samples [8], 26 to 33 pulmonary samples [8])

	1	2	3	4	5	6	7	8	9	10	11
Original	7.5	8.75	8.5	7.75	7	7	6.75	8.25	8	7.5	6.75
G.722	7	7.5	7.5	7.25	6.25	6.25	5.5	7.75	7	7	6.5
Stethoscope-Codec	6.5	7.5	7.25	7	6	6	5.75	7	7	6.75	6.5

	12	13	14	15	16	17	18	19	20	21	22
Original	6.25	7.5	**7.25**	8.25	6.5	7.5	8	6	5.75	5.5	6.75
G.722	6.25	7.25	6.75	7.75	6.25	7	7.75	5.75	5	4.75	6.25
Stethoscope-Codec	6	6.75	6.5	7.25	6	6.75	7.25	5.5	5	5	5.5
	23	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**
Original	6.75	6.5	5.5	5.25	6.75	7	6.25	6	6.75	5.75	6.25
G.722	6.25	6.25	6	5	5.5	6.25	5.75	5.75	6	5.75	6.25
Stethoscope-Codec	6.5	6.25	6	5.25	5.25	6	5.75	5.75	5.75	5.5	6

Table 4. Average grades per physician

	Original	**G.722**	**Stethoscope-Codec**
Dr. 1	7.33333333	7.42424242	7.90909091
Dr. 2	7	5.75757576	5.42424242
Dr. 3	6.72727273	6.24242424	5.87878788
Dr. 4	6.54545455	6.21212121	5.60606061
Average	**6.90151515**	**6.40909091**	**6.20454545**

6 Conclusion and Future Work

Despite the new Proposed Stethoscope-Codec provides better objective results than G.722 when comparing figures of merit such as RMSE and PRD, subjective medical assessment of the former is slightly lower. The new Proposed Stethoscope-Codec reduces the bit rate in a 4:1 proportion, going from 64 Kbps to 16 Kbps, and decreasing the RTF in a greater than 3:1 proportion, going from 0.02122 to 0.00588. The Proposed Stethoscope-Codec could be a good alternative to G.722 in cases where communication networks are precarious and with small bandwidth. Such is the case of most rural areas in developing countries. Our codec would therefore allow a higher number of auscultations at the same time, reducing the latency and requiring lower-processing-capacity computers thanks to the RTF reduction. The medical assessment presented should be taken with care, since the number of physicians consulted is still small. A more in-depth clinical study with patients suffering of a certain kind of pathology would be needed in order to validate the medical use of the Proposed Stethoscope-Codec.

References

1. Hedayioglu FL, Mattos SS, Moser L, de Lima ME. 2007. Development of a tele-stethoscope and its application in pediatric cardiology. in Indian J Exp Biol. 2007 Jan;45(1):86-92.
2. Sankaran, P. RMK Eng. Coll., Chennai, India Chandrasekaran, K.;Baig, A.H.; Moll, C.L. 2010. Development of a tele-stethoscope: Indian perspective, 2010, in Proc. IEEE/ICME International Conference on Complex Medical Engineering (CME), 2010.
3. Ignacio Foche Pérez: Desarrollo de un teleestetoscopio digital Bluetooth para zonas rurales aisladas de países en desarrollo, Proyecto Fin de Master en Redes de Telecomunicación para Países en Desarrollo, Escuela Técnica Superior de Ingeniería de Recomunicaciones, Universidad Rey Juan Carlos.
4. M. Abella and J. Formolo: Comparsion of the Acoustic Properties of Six Popular Stethoscopes, Department of Internal Medicine, St. John Hospital, Detroit, Michigan 48236.
5. I. Mazic, S. Sovilj and R. Magjarevic: Analysis od Respiratory Sounds in Asthmatic Infants, Measurement Science Review, Vol. 3, Section 2, 2003.
6. Alan V. Oppenhheim, Alans S. Willsky, S. Hamid Nawab: Señales y Sistemas, pp. 527-528, Segunda edición, Pearson Educación.
7. http://www.ekiga.org/, Ekiga Softphone.
8. http://www.med.umich.edu/lrc/psb/heartsounds/index.htm, University of Michigan Heart Sound and Murmur Library.
9. http://sox.sourceforge.net/, Sound eXchange.

Morphological Processing of a Dynamic Compressive Gammachirp Filterbank for Automatic Speech Recognition

Joyner Cadore, Carmen Peláez-Moreno, and Ascensión Gallardo-Antolín

Universidad Carlos III de Madrid, Escuela Politécnica Superior,
Avda. de la Universidad 30, 28911 Madrid, Spain
{jcadore,carmen,gallardo}@tsc.uc3m.es
http://gpm.tsc.uc3m.es/

Abstract. The Dynamic Compressive Gammachirp ([8]) is presented for producing auditory-inspired feature extraction in Automatic Speech Recognition. The proposed acoustic features combine spectral subtraction and two-dimensional non-linear filtering technique most usually employed for image processing: morphological filtering. These features have been proven to be more robust to noisy speech than those based on simpler auditory filterbanks like the classical mel-scaled triangular filterbank, the Gammatone filterbank and the passive Gammachirp in a noisy Isolet database.

Keywords: spectral subtraction, morphological processing, automatic speech recognition, auditory-based features.

1 Introduction

Machine performance in Speech Recognition tasks is still far from that of humans.

Among non-satisfactorily tackled challenges is background noise. One way to address these limitations is trying to imitate human acoustic capabilities, e.g. finding a more suitable auditory model. In this paper we focus on the noise problem where humans are known to perform remarkably well whilst machines still lag behind [13].

There are many solutions inspired by the Human Auditory System (HAS) aimed at solving this issue, e.g. feature extraction based on the well-known mel-frequency cepstral coefficients (MFCC), and on the Gammatone-based coefficients (GTC) are some of many examples. Other solutions use the so-called *spectro-temporal features*, that consider both the time and frequency domains in the feature extraction stage ([10]).

Following that line of work, the authors presented a morphological filtering over a cochleogram (i.e. an auditory spectrogram) of a noisy signal to mimic some properties of the HAS, such as frequency and temporal masking [3, 4]. Two types of basic cochleograms were compared: the first, based on classical

mel-scaled critical bands and the second, on gammatone auditory filters presenting comparable results. In this paper, we propose the use of more detailed formulation of the auditory filtering procedure based on successive improvements of the gammatone formulation by Irino and Patterson: the dynamic compressive gammachirp (dcGC) auditory filterbank [8].

Gammatones (GT) have been already employed in a number of papers for ASR ([15] among others) while passive gammachirp (pGC) of [8] is less widespread ([9]). However, to our knowledge, only preliminary experiments on syllable recognition have been deployed for dcGC with the purpose of showing the scale-shift covariance properties of this auditory model that provides a better way for adapting ASR acoustic models to vocal tract length variations ([12]). A different approach for introducing dynamic auditory filtering is based on the application of the Dyn non-linear operator applied on the compressive gammachirp (cGC) [5] demonstrating robustness improvements over the conventional MFCC in several noisy conditions on TIMIT.

In this paper we have found that dcGC outperforms cGC in noisy mismatched conditions specially in combination with the well-known Spectral Subtraction (SS) preprocessing method. The morphological postfiltering applied to the resulting cochleogram also provides enhanced performance (though marginally). The new resulting features have been tried on a hybrid MLP/HMM recognizer.

This paper is organized as follows: in section 2 we describe the alternative methods for auditory filtering aforementioned. Section 3 presents the pre and postprocessing stages of our feature extraction method. Section 4 describes the experiments and results obtained to end with the conclusions and futher work in section 5.

2 Gammachirp auditory filters

The Gammachirp auditory filter [8] is an extension of the Gammatone filter. The impulse response of a gammachirp filter is defined by:

$$g(t) = kt^{n-1}exp(-2\pi b\mathrm{ERB}(f_c)t) \times exp(j2\pi f_c t + jc\ln(t) + j\phi) \quad (1)$$

where n is the order of the filter, k defines the output gain, b defines the envelope of the gamma distribution, c is the chirp factor, f_c is the filter's central frequency, ϕ is the phase and ERB is the Equivalent Rectangular Bandwidth defined in [11]. When $c = 0$, eq. (1) reduces to the impulse response of the Gammatone filter.

The Fourier magnitude spectrum of the Gammachirp filter is:

$$\left|G_C(f)\right| = A \cdot \left|G_T(f)\right| \cdot exp(c\theta(f)) \quad (2)$$

$$\theta(f) = arctan(\frac{f - f_c}{b\mathrm{ERB}(f_c)}) \quad (3)$$

where $|G_T(f)|$ is the Fourier magnitude spectrum of the Gammatone filter.

From (2), (3) it is possible to obtain the next three types of Gammachirp filters [8]:

- The Passive Gammachirp (pGC): level-independent and representing the passive basilar membrane.
- The Compressive Gammachirp (cGC): level-dependent and simulating the active mechanism in the cochlea.
- The Dynamic Compressive Gammachirp (dcGC): including a fast-acting level control circuit for the cGC filter, two-tone suppression and compression.

3 Spectral Subtraction and Morphological Filtering

The filterbanks described in section 2, furthermore the very well-known Mel-scaled triangular filterbank and Gammatone filterbank, have been embedded in a feature extraction system that includes spectral subtraction as a preprocessing stage and a morphological filtering as postprocessing aimed at imitating both temporal and instantaneous masking in the HAS.

3.1 Spectral Subtraction

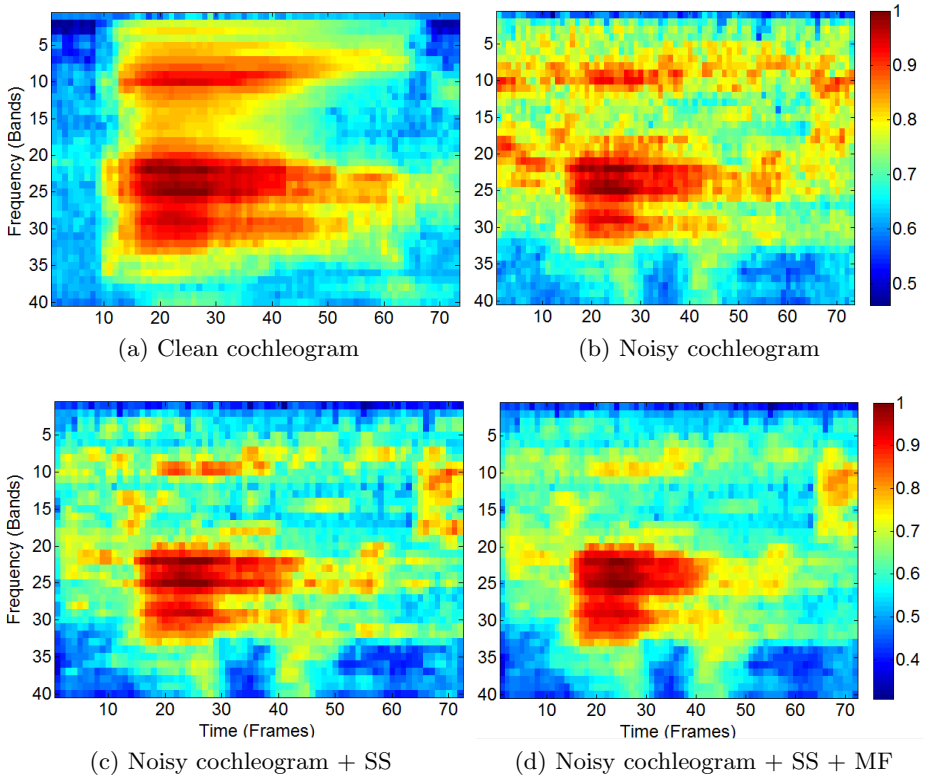
Spectral Subtraction (SS) is a classical procedure to remove noise from speech [1]. Figures 1a and 1b shows the cochleogram of a clean and noisy speech sample, respectively. Spectral subtraction obtains an estimate of the density spectrum of the noise and performs a subtraction in the frequency domain. The SS was applied in the magnitude domain. Fig. 1c depicts the resulting cochleogram of the so cleaned signal.

3.2 Morphological Filtering of cochleograms

Auditory masking has been largely studied as regarding the influence of some frequencies on others simultaneously present in the spectrum, or *simultaneous masking*, or as regarding the influence of the same frequencies at different time instants, or *temporal masking*. Therefore its effects can be observed both in time and frequency domains requiring a two dimensional representation to jointly consider the two of them.

The application of an auditory motivated filterbank (as those described in section 2) produces a more uniform representation of the simultaneous masking effects that is certainly more amenable for a computational modelling since the asymmetrical masking threshold becomes *almost* independent of the scaled frequency. An auditory spectrogram (sometimes referred as *cochleogram*) substitutes the usual linear spectral representation by auditory motivated filterbanks uniformly distributed in a scaled frequency.

On the other hand, Mathematical Morphology (MM) is a theory for the analysis of spatial structures [14] whose main application domain is in image processing as a tool for thinning, pruning, structure enhancement, object marking, segmentation and noise filtering. Thus, Morphological Filtering (MF) becomes an adequate operation to consider both domains of auditory masking.



(a) Clean cochleogram (b) Noisy cochleogram

(c) Noisy cochleogram + SS (d) Noisy cochleogram + SS + MF

Fig. 1: Resulting auditory cochleograms of a utterance with added babble noise at 5dB.

3.3 Morphological operations

We use S with implicit frequency band index n and temporal frame k to represent a cochleogram $S(n,k)$. Then erosion and dilation with a given mask M can be represented as matrix operations:

$$S \ominus M = \{p \in \mathbb{R}^2 \mid p = m - s, m \in M,\ s \in S\} \tag{4}$$

$$S \oplus M = \{p \in \mathbb{R}^2 \mid p = s + m, s \in S, m \in M\} \tag{5}$$

Erosion is used to shrink or reduce objects, while dilation, being the dual to erosion, produces an enlargement. Both are irreversible.

Opening and closing are used to remove small objects in images, typically noise, their behaviour with respect to, for instance, salt and pepper noise, being dual to each other. The composite opening and closing operation adopt the following form:

$$S \circ M = (S \ominus M) \oplus M \tag{6}$$

$$S \bullet M = (S \oplus M) \ominus M \tag{7}$$

We use the opening operator over the cochleogram with the discrete representation of the mask depicted in fig. 3 to obtain:

$$S'' = S' + S' \circ M \tag{8}$$

The resulting cochleogram S'' of a sample noisy signal is shown in fig. 1d. Finally an DCT is performed to obtain $\hat{S}$ as shown in fig. 2.

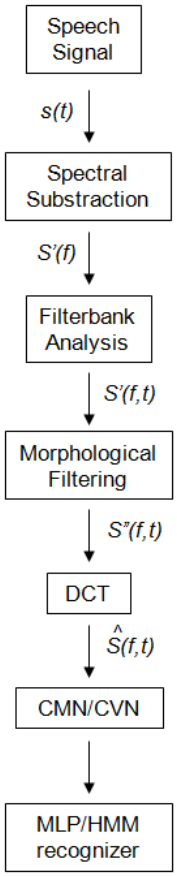


Fig. 2: Block diagram of the ASR system.

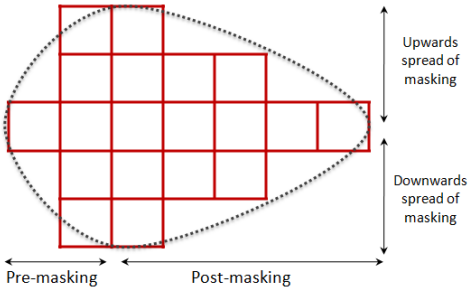


Fig. 3: Auditory-motivated mask (dotted line) and its discretization (solid line). The horizontal and the vertical axis represents time and frequency scales, respectively.

4 Experiments

4.1 General Description and Feature Extraction

The block diagram of the ASR system used in the experimentation is depicted in fig. 2.

Four different types of auditory filterbanks are considered: triangular mel-scaled, Gammatone, passive Gammachirp and dynamic compressive Gammachirp filters, yielding Mel-Frequency Cepstrum Coefficients (MFCC), Gammatone-based (GTC), passive Gammachirp-based (pGC) and dynamic compressive Gammachirp-based (dcGC) features, respectively.

In all the cases, speech is analyzed using a 25 ms window every 10 ms. The corresponding auditory filterbank is composed of 40 bands. After the DCT, coefficients C_0 to C_{12} are kept. Adding their corresponding delta and acceleration coefficients compose 39 dimensional vectors. The last step in the feature extraction stage was applying mean and variance normalization on either type of coefficient.

4.2 ISOLET Testbed

In the experimentation, we use the ISOLET testbed [7]. ISOLET is a database of letters of the English alphabet spoken in isolation. The database consists of 7800 spoken letters (two productions of each letter pronounced by 150 different speakers). Specifically, we use a version called Noisy-ISOLET: the speech signals of ISOLET plus 8 different noise types at different SNRs (clean, 0dB, 5dB, 10dB, 15dB and 20dB).

The experiments using the ISOLET testbed are performed over an hybrid MLP/HMM ASR system [2]. A context of 5 frames is used yielding an input of 195 elements to the MLP. The hybrid MLP/HMM system is tested in two different conditions: *mismatched*, where the system is trained using clean speech and *matched* where the training set is composed of a balanced combination of

speech contaminated with the different noises of the database at several SNRs. A 5-fold cross-correlation procedure has been employed to improve statistical significance [6].

4.3 Results

Table 1 summarizes the experiments performed to study the impact of the Spectral Subtraction (SS), the Auditory filtering and the Morphological Filtering (MF).

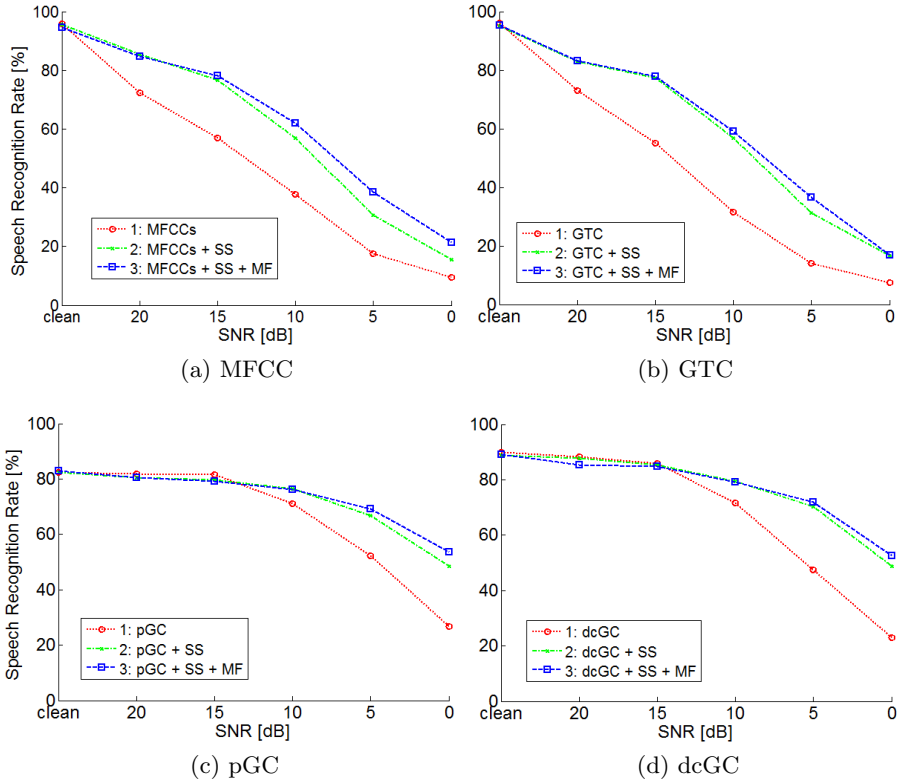
Table 1: Recognition results in terms of *WER* [%] and 95% confidence intervals.

Features	Mismatched	Matched
MFCC	51.80 ± 1.24	16.45 ± 0.92
MFCC + SS	40.85 ± 1.22	16.95 ± 0.93
MFCC + SS + MF	37.03 ± 1.20	17.05 ± 0.93
GTC	53.78 ± 1.24	17.15 ± 0.94
GTC + SS	40.28 ± 1.22	16.95 ± 0.93
GTC + SS + MF	38.50 ± 1.21	16.85 ± 0.93
pGC	34.03 ± 1.18	25.78 ± 1.09
pGC + SS	27.60 ± 1.11	26.63 ± 1.10
pGC + SS + MF	26.45 ± 1.09	27.95 ± 1.11
dcGC	32.40 ± 1.16	20.08 ± 0.99
dcGC + SS	23.25 ± 1.05	20.38 ± 1.00
dcGC + SS + MF	22.98 ± 1.04	20.18 ± 1.00

As for the *mismatched* condition when the auditory filterbank is considered alone, pGC and dcGC features achieve better results than MFCC and GTC, being the differences statistically significant. This fact shows the robustness of pGC and dcGC parameters in noisy conditions. The use of SS clearly improves the corresponding baselines where the best results are obtained with the dcGC + SS features, being the differences statistically significant again with respect. The sequential use of both techniques (SS + MF) improves the recognition rates of the system compared to the baseline and SS cases, specially for the MFCC parameterization. Nevertheless, dcGC + SS + MF achieves the best performance with a relative error reduction of around 38% with respect to MFCC + SS + MF.

For the *matched* condition, best results are obtained with MFCC and GTC features. Besides, since no significant improvements are achieved by using SS or SS + MF the aforementioned techniques seem to be more suitable for the *mismatched* case.

Figure 4 shows the Recognition Rates achieved by the different techniques as a function of the SNR for the *mismatched* condition. As can be observed,



(a) MFCC (b) GTC

(c) pGC (d) dcGC

Fig. 4: SRR [%] vs SNR. Mismatched case.

the behaviour of MFCC and GTC features is rather different to the one shown by pGC and dcGC. The performance of the two first types of features degrades significantly for all SNRs, whereas the recognition rates achieved by pGC and dcGC only decrease drastically for the lower SNRs (0 dB - 10 dB). The combination of SS or SS + MF techniques with MFCC and GTC features improves the recognition rates for all SNRs. However, for pGC and dcGC, SS or SS + MF is only effective for SNRs below 15 dB. In any case, although in clean conditions, MFCC and GTC (alone or with SS or SS + MF) report the best results, the corresponding systems based on pGC and dcGC features achieve similar results for high SNRs and clearly outperform MFCC and GTC for low SNRs. Finally, when comparing pGC and dcGC, it can be observed that both sets of features exhibit similar trends, but dcGC is better than pGC for clean and high SNRs conditions.

5 Conclusions and future work

The Dynamic Compressive Gammachirp have been employed for producing auditory-inspired feature extraction in Automatic Speech Recognition. The proposed acoustic features also combine spectral subtraction and two-dimensional non-linear filtering technique most usually employed for image processing: morphological filtering. These features have been proven to be more robust to noisy speech than those based on simpler auditory filterbanks like the classical mel-scaled triangular filterbank, the Gammatone filterbank and the passive Gammachirp in a noisy Isolet database.

Future lines of work include testing different types of masks and morphological operations and further study the synergies between the effects of the mask and the dcGC-based features.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Science and Innovation CICYT Projects No. TEC2008-06382/TEC and No. TEC2011-26807. We would like to thank Prof. Toshio Irino for kindly providing the software for Gammachirp auditory filters.

References

1. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79. vol. 4, pp. 208–211. IEEE (1979)
2. Bourlard, H., Morgan, N.: Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. Adaptive Processing of Sequences and Data Structures pp. 389–417 (1998)
3. Cadore, J., Gallardo-Antolín, A., Peláez-Moreno, C.: Morphological processing of spectrograms for speech enhancement. Advances in Nonlinear Speech Processing pp. 224–231 (2011)
4. Cadore, J., Valverde-Albacete, F.J., Gallardo-Antolín, A., Peláez-Moreno, C.: Auditory-inspired morphological processing of speech spectrograms. Cognitive Computation (Jul 2012 (accepted))
5. Gauci, O., Debono, C., Micallef, P.: A nonlinear feature extraction method for phoneme recognition. In: Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean. pp. 811–815 (2008)
6. Geisser, S.: The predictive sample reuse method with applications. Journal of the American Statistical Association 70, 320–328 (1975)
7. Gelbart, D., W., H., Holmberg, M., Morgan, N.: Noisy ISOLET and ISOLET testbeds (Feb 2011), `http://www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet/`
8. Irino, T., Patterson, R.D.: A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE Transactions on Audio, Speech, and Language Processing 14(6), 2222–2232 (Oct 2006)
9. Maganti, H., Matassoni, M.: A Level-Dependent Auditory Filter-Bank for Speech Recognition in Reverberant Environments. Twelfth Annual Conference of the International Speech Communication Association (2011)

10. Meyer, B., Kollmeier, B.: Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Communication (53), 753–767 (2010)
11. Moore, B., Glasberg, B.: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America 74, 750 (1983)
12. Patterson, R., Walters, T., Monaghan, J., Feldbauer, C., Irino, T.: Auditory speech processing for scale-shift covariance and its evaluation in automatic speech recognition. Audio, Transactions of the IRE Professional Group on pp. 3813–3816 (Dec 2009)
13. Peláez-Moreno, C., García-Moral, A., Valverde-Albacete, F.: Analyzing phonetic confusions using formal concept analysis. The Journal of the Acoustical Society of America 128(3), 1377–1390 (2010)
14. Serra, J., Soille, P. (eds.): Mathematical Morphology and its Application to Image Processing. Computational Imaging and Vision, Kluwer Academic (1994)
15. Yin, H., Hohmann, V., Nadeu, C.: Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency. Speech Communication (53), 707–715 (2010)

Manipulation of Unvoiced Fricatives for Improved Discrimination by Hearing Impaired

Rami Cohen[1], Oren Ierushalmi[1], Yizhar Lavner[1,2], Michal Genussov[1], and Ami Steiner[3]

[1] Signal and Image Processing Lab (SIPL), Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
rc@tx.technion.ac.il,orenierushalmi@gmail.com,michalgenus@gmail.com
[2] Department of Computer Science, Tel-Hai College
Upper Galilee, 12210, Israel
yizhar.lavner@gmail.com
[3] Ami Steiner Technologies
Tel Aviv, 64375, Israel
ami.steiner@gmail.com

Abstract. Many hearing impaired people suffer from hearing loss in the high-frequency range. This impairment leads to a confusion between unvoiced fricatives such as /s/ and /sh/ (aka sibilants) that current hearing instruments find difficult to overcome. In this paper we develop an algorithm for improving the discrimination between these sibilants. Practical methods for modifying the spectral characteristics of these phonemes by manipulating their Linear Predictive Coding coefficients are introduced. Results of experiments conducted on hearing impaired subjects, for evaluating the proposed methods, are provided, and show their potential.

Keywords: Hearing impaired, Linear Predictive Coding, Unvoiced fricatives, Sibilants

1 Introduction

According to comprehensive statistics of the World Health Organization [1], hearing impairment is the most frequent sensory deficit in human populations, affecting more than 250 million people in the world. Consequences of hearing impairment include inability to interpret speech sounds, resulting in a reduced ability to communicate and delay in language acquisition, among several economic, social and educational disadvantages.

Special difficulties appear in hearing frequencies beyond 2.5kHz. This may cause problems in the perception of various phonemes in conversational speech, typically in discriminating between phonemes that are categorized as unvoiced fricatives, such as /s/, /sh/, /f/ and /th/. Current hearing instruments can amplify different ranges of frequencies and overcome background noise, but usually there is no attempt at improving the perception and discrimination of individual phonemes.

There are several options that can be considered for improving the audibility of unvoiced fricatives. One option is to extend artificially their time duration using time-scale modification methods. Another possibility is to replace each of the relevant phonemes by a pre-recorded one that sounds better or by a unique vocal code [2]. An obvious disadvantage of these methods is that they can severely affect the natural flow of speech.

In previous studies that attempted to enhance the distinction between pairs of phonemes that are confused by hearing impaired, different cues were used and manipulated, such as the consonant-vowel intensity ratio [3], vowel duration [4], amplification of burst/murmur cues for final stop voicing discrimination [5] or transposition of high frequency information to lower frequencies [6].

In another study [7], log-area ratio coefficients derived from the Linear Predictive Coding (LPC) coefficients were used to produce signals that are acoustically and perceptually in between two original tokens such as /l/ and /r/. When extrapolation along the line connecting the natural tokens in model coefficients space was applied, "exaggerated" phonemes were produced. These phonemes were found to be more distinctive for people who lack the ability to discriminate the natural phonemes.

In this paper we propose a method for exaggerating the distinction between the unvoiced fricatives /s/ and /sh/, by modifying their spectra using manipulations of their LPC all-pole model: pole rotation to modify the frequency of spectral peaks, pole addition to increase their bandwidth, or a combination of the two. Our method is based on the observation that sibilant fricatives can be distinguished according to the location of their first (lowest) spectral peak, which is around 4 kHz for /s/ and 2.5kHz for /sh/ [8].

In order to evaluate the performance of the proposed manipulations, experiments involving hearing impaired people were conducted. Two sets of tests, one based on Just Noticeable Difference (JND) and the other on an adaptive staircase variation of intensity levels with nonsense syllables (logatoms) [9], which determines the recognition threshold for nonsense syllables such as 'asa', 'asha', etc., were conducted, and both will be described later.

This paper is organized as follows. In section 2, a short review of the LPC model is presented and the manipulation methods are introduced. In section 3, the results of experiments performed for the evaluation of our methods are presented. Finally, conclusions are drawn in section 4.

2 Algorithm Description

In this section we introduce several methods for improving the distinction between unvoiced fricatives, specifically /s/ and /sh/. The phonemes used in our work (10 instances for each phoneme) were extracted from the TIMIT database [10], sampled at a rate of 16kHz.

2.1 LPC Coefficients

The linear prediction model we use here is a discrete-time model of speech production, based on the source-filter model [11]. In the linear prediction model, the vocal tract is represented by an all-pole filter [12]. This filter is excited by either an impulse train at pitch periodicity that leads to voiced speech, or an input of white noise that produces unvoiced speech. It is assumed that the current sample $s(n)$ is a linear combination of the p previous samples, with a possible residual (error) $e(n)$:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n) \tag{1}$$

where a_k (the *LPC coefficients*) are obtained by minimizing the expectation of the squared error $e^2(n)$, using iterative methods, such as the Levinson-Durbin algorithm [13]. From the last equation we get the following relation in the frequency domain:

$$S\left(e^{j\omega}\right) = E\left(e^{j\omega}\right) V\left(e^{j\omega}\right) \tag{2}$$

where $E\left(e^{j\omega}\right)$ is the Fourier transform of the residual $e(n)$, and $V\left(e^{j\omega}\right)$ (known as the all-pole filter) is defined as:

$$V\left(e^{j\omega}\right) \triangleq \frac{1}{1 - \sum_{k=1}^{p} a_k e^{-jk\omega}} \tag{3}$$

and is used as a short-time spectrum estimation of the speech signal, but without (if p is not too big) the random fluctuations that stem from the excitation - i.e., the spectral envelope. Therefore, we focused on manipulating the poles of the all-pole filter in order to modify the spectral content of the phonemes, aiming at improving the distinction between unvoiced fricatives.

2.2 Methods for Spectral Modification

Three methods, which are based on manipulating the poles of a 3$^{\text{rd}}$ order LPC filter of the signals, were used for modifying the phonemes' spectra. The choice of this order enables the manipulation of the main peak of the spectral envelope. It should be noted that these manipulations require a preliminary step of unvoiced fricatives spotting [16].

The goals of these methods were twofold: 1) to emphasize the phonemes so that their inherent properties would be more prominent, and 2) to modify the phonemes so that they will be more distinguishable by the hearing impaired. This can be achieved by modifying the peaks of the spectral envelope using the poles of the linear prediction analysis.

For this aim, pole rotation, pole addition, or both were carried out. It should be emphasized that the energy of each of the modified phonemes was matched to the original one, in order to avoid effects of amplification of the phonemes. The rest of this section is devoted to describing these methods.

Pole rotation In this method, the complex poles of the LPC filter of the tested phoneme were rotated towards the poles of an exaggerated reference phoneme. The latter phoneme was selected in advance, as the most distinctive one out of the phonemes in our database, and then was exaggerated using the following procedure.

First, the complex conjugate poles of the reference phoneme were rotated in order to exaggerate the discrimination between /s/ and /sh/. This has been done by rotating the LPC poles towards a higher frequency in the case of /s/, or towards a lower frequency in the case of /sh/. A value of $\pi/8$ (1000Hz) was chosen experimentally for this frequency shift.

Then, the complex poles of the phoneme to be emphasized were rotated to the frequencies of the exaggerated reference phoneme, obtaining a new LPC filter. Finally, the tested phoneme's residual signal was filtered by the new linear prediction filter, resulting in an emphasized phoneme.

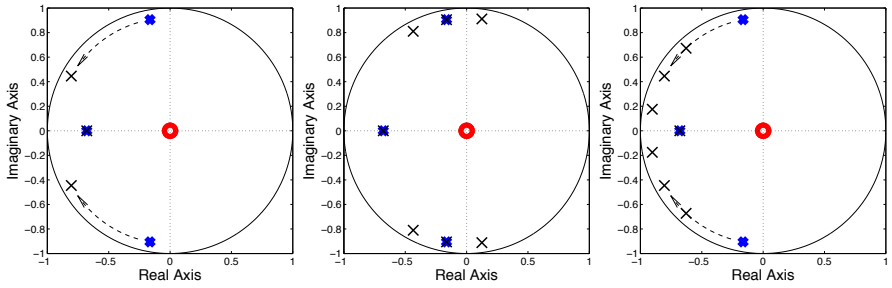
An example (for /s/) of the pole-zero map after this manipulation and the corresponding frequency response of the LPC filter, is shown in Figures 1a and 2a, respectively. It can be seen that the peak of /s/ is moved to a higher frequency. This produces a pair of /s/ and /sh/ that are acoustically more distinctive, since the spectral peak of /s/ is modified to a higher value relative to its original value (around 4kHz), while that of /sh/ is changed to a lower value (relative to its original value which is around 2.5 kHz) after pole rotation. Therefore, the phonemes are aimed to be perceptually more distinct after this manipulation.

Pole addition In this method, the bandwidth of the main peak of the phoneme's spectral envelope is expanded, by adding an extra pole to both sides of each complex pole. This is achieved by adding a pole with a radius of the original complex poles at an angle difference of $\pi/10$ (800Hz) from both sides of each complex pole. A new filter was obtained using the new set of poles, and the emphasized phoneme was obtained by filtering the original residual by this new filter.

An example (for /s/) of the pole-zero map after this manipulation and the corresponding frequency response of the LPC filter, is shown in Figures 1b and 2b, respectively. It can be seen that the peak of the manipulated /s/ is wider, i.e., the bandwidth of this peak is larger than the original one. It should be noted that this method does not require the classification of the specific unvoiced fricative beforehand, and as such is easier to implement in hearing aids.

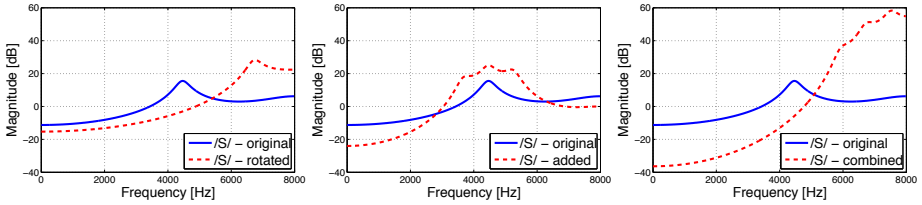
Combined methods In this method, the two previous methods are simply combined. That is, the poles of the tested phoneme are rotated, and then two poles are added near each of the resulting complex poles, as described above.

An example (for /s/) of the pole-zero map after this manipulation and the corresponding frequency response of the LPC filter, is shown in Figures 1c and 2c, respectively. Evidently, both effects are achieved by this method: the peak is moved to a a higher frequency, and the bandwidth around this frequency is increased.



(a) /s/ - Pole rotation (b) /s/ - Pole addition (c) /s/ - Combined methods

Fig. 1: Manipulation methods: Pole-zero map of the LPC filter (Bold 'x' denotes the original poles)



(a) /s/ - Pole rotation (b) /s/ - Pole addition (c) /s/ - Combined methods

Fig. 2: Manipulation methods: LPC filter frequency response

3 Performance Evaluation

We used two tests in order to evaluate the performance of the methods described above, employing two different procedures: A test based on Just Noticeable Difference (JND) and a test with an up-down staircase procedure of intensity levels in which logatom syllables were used. The tests involved 8 and 6 hearing impaired people, respectively, all of them with moderate hearing degradation in the high-frequency range. All subjects were paid for their participation. The tests were conducted in a quiet room using high-quality headphones. Each test session lasted about 30 minutes.

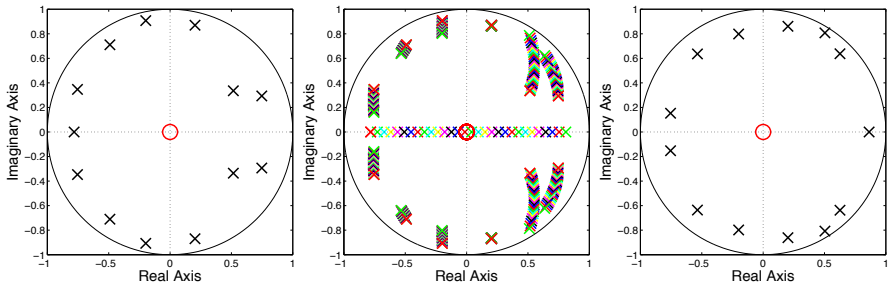
3.1 JND Test

Just Noticeable Difference (JND) is the minimal difference in stimulus intensity or level that leads to a detectable change in perceptual experience. In this test, we measure the number of JND steps in a succession of stimuli (phonemes), each of which with a small difference of a certain attribute relative to its predecessor.

Our assumption is that this number can be used as an estimate of the perceived discrimination between a pair of phonemes, where a large number of steps

indicates that a pair of phonemes can be better distinguished. Thus, to measure the perceptual discrimination between two phonemes, a base phoneme /x/ and a final phoneme /y/, a succession of a pre-defined number N of intermediate phonemes ('/z_i/', $i = 1, 2, ..., N$) is produced.

The stimuli are produced by rotating the LPC poles of order 13 of /x/ towards /y/. This order was selected to achieve high detailed representation of the spectra of both base and target phonemes. A logarithmic scale was used, where the angular difference is small for angles corresponding to lower frequencies and large for angles corresponding to higher frequencies, in accordance with the logarithmic sensitivity of the auditory system. The radii of the poles were modified as well during this process, using a linear scale. This process is depicted in Figure 3.



(a) The pole-zero map of the base phoneme /s/ (b) Transition process (c) The pole-zero map of the final phoneme /sh/

Fig. 3: JND transition process: example

At each stage, the residual signal is filtered by a new filter, obtained by the modified poles, in order to produce a modified (intermediate) phoneme /z_i/. The residual for each intermediate phoneme is chosen as the weighted sum of the two residual signals: the residual of the base phoneme /x/ and that of the final phoneme /y/. The weight is set as the relative distance between the angles of the complex poles of /z_i/ and the corresponding poles of /x/ and /y/, so that the intermediate residual contains characteristics of both residuals.

At each stage of the test, the subject hears two words: 'a/x/a' (where /x/ is the base phoneme) and 'a/z_i/a' (where /z_i/ is the current intermediate phoneme, beginning with $i = 1$). The subject is asked whether the two words sound the same or not. If there is a difference, it means that there is a *JND step* between the phonemes /x/ and /z_i/. In this case, the base word is replaced by the intermediate 'a/z_i/a', and the test continues by comparing it to the 'a/z_{i+1}/a' of the next logarithmic step. If the phonemes are perceived as same, the test continues with the same 'a/x/a' and with 'a/z_{i+1}/a'. Subjects can hear the stimuli repeatedly at will.

The test ends when a final word, 'a/y/a', is reached. At this stage, the total number of JND steps is counted. A schematic description of the JND procedure is shown in Figure 4.

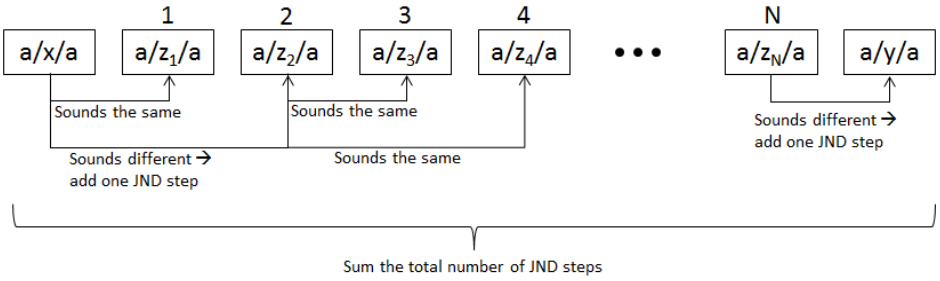


Fig. 4: An example of the JND procedure. N intermediate phonemes are produced between the base phoneme and the final phoneme.

For our purposes, the base phoneme /x/ and the final phoneme /y/ were chosen randomly as either /s/ or /sh/. Each test was performed twice - first for the original /s/ and /sh/, and then for the manipulated ones (for each type of manipulation). The order of the tests (i.e., comparing original or manipulated phonemes) was chosen randomly. The result of the test is defined as the number of times that the subject reported a difference between the words, namely, count of JND steps.

The JND test was conducted using $N = 15$ steps between /s/ and /sh/. Eight subjects (5 males and 3 females, of various ages) participated in this test. The results are summarized in Figure 5, where the number of subjects with increased number of JND steps ("improved"), decreased number ("worse") or with the same number of JND steps ("no change"), compared to the corresponding number for the original phonemes, is shown for each manipulation method. According to these results, it can be seen that the pole addition method and the combined method have better performance over the original phonemes.

3.2 Intensity level threshold estimation using adaptive staircase procedure

In order to evaluate new hearing instrument algorithms and/or adaptive hearing instrument processes, a discrimination measurement method, which is sufficiently sensitive for cases of mild to moderate hearing loss, should be used. The aim of this test is to determine the recognition intensity threshold (in dB) for nonsense syllables (denoted as logatoms) such as 'asa', 'asha', etc. [9].

In a preliminary calibration stage the subject heard an English sentence with different intensities. Initially, the intensity of the sentence was decreased in steps of 5dB, until a hearing level at which the subject reported that the sentence could

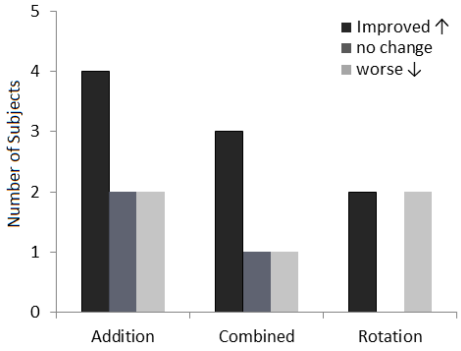


Fig. 5: JND results

not be heard was reached. This level was defined as the *threshold level*. Then, in the next trials, the intensity of the sentence was increased (again, in steps of 5dB) until a hearing level was reached where listening to the sentence was most comfortable. This level was defined as the *most comfortable level*.

After the calibration stage, the subject heard in each trial a word from a set of VCV (Vowel-Consonant-Vowel) logatoms: 'asa', 'asha', 'afa' and 'atha'. In each trial, one logatom was randomly selected from a set of 10 stimuli, which consisted of the original /s/ and /sh/, the manipulated ones (by the three methods described in Section 2.2), and two distracters: /f/ and /th/.

After each trial, the subject reported which word from the set was heard. The subject could also choose to repeat the trial, or select the option 'can't hear'. An adaptive staircase procedure [14,15] was administered by an interactive computer program which also recorded the listener responses.

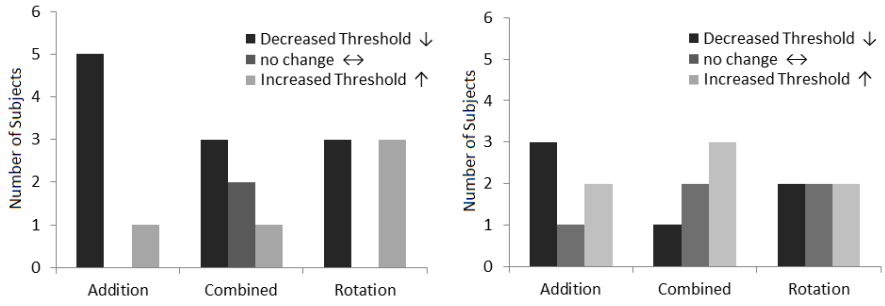
The staircase procedure was conducted as follows. The intensity level, which was initialized as the most comfortable level, was decreased after each correct answer, and was increased to the previous intensity level for a wrong answer. Each decreasing step was half of the previous decreasing steps.

The test continued until a predetermined maximal number of trials was reached, or after detecting 5 reversals of the intensity for each phoneme. The score of the test was determined as the average of the intensities in last 4 trials. For each subject, the final score was considered as the average of the scores for each phoneme. A maximal number of 20 trials was used. We expected to find lower scores for manipulated phonemes, i.e., the latter would be better perceived at lower intensities compared to the original phonemes.

Six subjects participated in the threshold estimation test (5 males, 1 female, of various ages). The results are depicted in Figure 6, where the number of subjects with decreased/increased recognition threshold, compared to the corresponding threshold of the original phonemes, is shown.

According to these results, it can be seen that in the case of /s/ (Figure 6a), the method of pole addition (improvement for 5 subjects out of 6), and to a

lesser extent the combined method, have better performance over the original phonemes. For /sh/ (Figure 6b), there is an improvement only in the case of pole addition.



(a) /s/ (b) /sh/

Fig. 6: Threshold estimation results

4 Conclusion

In this work, we introduced a method for exaggerating unvoiced fricatives phonemes by manipulating their poles, in the framework of linear prediction all-pole model. The evaluation tests show that for some of the modifications, and especially for pole addition, the pair of manipulated phonemes /s/ and /sh/ were more distinguishable for hearing impaired subjects compared with the original pair of phonemes, indicating the potential of using the method to enhance the distinction between the phonemes.

There are many types of hearing impairments. Focusing on a limited range of impairments may help to achieve more significant results. It is likely that every impairment type requires a different approach.

In addition, many subjects claimed that their major difficulty in understanding speech correctly in their daily lives is the presence of noise from the surroundings, such as other peoples talking. Thus, a model for noise may be used for further testing the above methods for enhancing the distinction between phonemes in the presence of different levels of noise.

This study lays the foundations for a broader research on the exaggeration of unvoiced fricatives (rather than just /s/ and /sh/). Furthermore, each language has a different set of unvoiced fricatives that should be treated accordingly. An ongoing research is currently performed in our lab, in which additional phonemes are examined, as well as additional manipulation methods.

Acknowledgement

The authors would like to thank Phonak AG (Switzerland) for their assistance. The support of the staff of the Signal and Image Processing Lab (SIPL) is greatly appreciated.

References

1. Mathers C., Smith A., Concha M.: Global burden of hearing loss in the year 2000. Technical report, World Health Organisation, Geneva (2007)
2. Bauer D., Plinge A., Finke M.: Selective Phoneme Spotting for Realization of an /s, z, C, t/ Transposer. In: 8th International Conference on Computers Helping People with Special Needs, pp. 153–161. Springer-Verlag, London (2002)
3. Shobha N. H., Thomas T. G., Subbarao K.: Influence of Consonant-Vowel Intensity Ratio on Speech Perception for Hearing Impaired Listeners. In: 2008 Congress on Image and Signal Processing, pp. 342–346. IEEE Press, Washington (2008)
4. Revoile S. G., Holden-Pitt L., Picket J. M., Brandt F.: Speech Cue Enhancement for the Hearing Impaired: I. Altered Vowel Durations for Perception of Final Fricative Voicing. Journal of Speech and Hearing Research 29, pp. 240–255 (1986)
5. Revoile S. G., Holden-Pitt L., Edward D., Picket J. M., Brandt F.: Speech-cue enhancement for the hearing impaired: amplification of burst/murmur cues for improved perception of final stop voicing. Journal of Rehabilitation Research and Development 24, pp. 207–216 (1987)
6. Robinson J. D., Baer T., Moore B. C. J.: Using transposition to improve consonant discrimination and detection for listeners with severe high-frequency hearing loss. International Journal of Audiology 46, pp. 293–308 (2007)
7. Protopapas A.: Modified LPC resynthesis for controlling speech stimulus discriminability. The Journal of the Acoustical Society of America 104, pp. 1855–1855 (1998)
8. Raphael L. J., Borden G. J., Harris K. S.: Speech Science Primer: Physiology, Acoustics, and Perception of Speech. Williams and Wilkins, Philadelphia (2006)
9. Boretzki M.: Structure and Evaluation of an Adaptive Logatom Test for Measuring the Identification of Consonants. Technical report, Phonak corporation, Switzerland (2009)
10. Garofolo J. S.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Technical report, Linguistic Data Consortium, Philadelphia (1993)
11. Rabiner L. and Schafer R.: Theory and Applications of Digital Speech Processing. Pearson, New Jersey (2011)
12. Deller J. R., Hansen J. H. L., Proakis J. G.: Discrete Time Processing of Speech Signals. IEEE Press, New York (1999)
13. Ljung L.: System Identification: Theory for the User. pp. 278–280, Prentice-Hall, New Jersey (1987)
14. Leek M. R.: Adaptive procedures in psychophysical research. Perception and Psychophysics 63, pp. 1279–1292 (2001)
15. Levitt H.: Transformed Up-Down Methods in Psychoacoustics. The Journal of the Acoustical Society of America 49, pp. 467–477 (2001)
16. Ruinskiy D., Dadush N. and Lavner Y.: Spectral and Textural Feature-Based System for Automatic Detection of Fricatives and Affricates. In: IEEEI 2010, pp. 771–775, Eilat, Israel (2010)

Multiclass Pitch Accent Classification for Assisting Manual Prosodic Labeling*

David Escudero-Mancebo[1], Francisco Vizcaíno-Ortega[2], César González-Ferreras[1], Carlos Vivaracho-Pascual[1], Mercedes Cabrera-Abreu[2], Eva Estebas-Vilaplana[3], and Valentín Cardeñoso-Payo[1]

[1] Department of Computer Science, Universidad de Valladolid, Valladolid, Spain
{descuder,cesargf,cevp,valen}@infor.uva.es
[2] Department of Modern Languages, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
{fvizcaino,mcabrera}@dfm.ulpgc.es
[3] Department of Modern Languages, Universidad Nacional de Educación a Distancia, Madrid, Spain
eestebas@flog.uned.es

Abstract. In this paper, we present an experiment on computer assisted prosodic labeling in which a labeler team validates or corrects ToBI pitch accents automatically predicted by a classifier. The innovative aspect of this automatic system is that it is not a deterministic prediction model, it offers the human transcriber more than one label per word (multi-class classification) and it is the transcriber who must decide on their appropriateness. The results make it evident that the procedure is indeed effective, since it not only reduces manual transcription time, but also improves global inter-transcriber consistency.

Keywords: Computer Assisted Prosodic Labeling, ToBI Labeling, Automatic Prosodic Labeling

1 Introduction

Prosodic labeling is a useful tool which provides relevant information for many applications. One such piece of information is the emphasis or accent whereby some parts of the utterance are highlighted. Knowing what parts of the utterance are accented helps improve various speech technology applications. Thus, detecting the prominent syllable in a word can be used to perform lexical disambiguation in Automatic Speech Recognition. In Dialog System, the knowledge of the emphatic words can serve to interpret or classify a message from either a semantic or a pragmatic viewpoint. As for Text-to-Speech synthesis, the correct identification between the prosodic form and the prosodic function is essential

* This work has been partially supported by Ministerio de Ciencia e Innovacion, Spanish Government (Glissando projects FFI2008-04982-C03-02 and FFI2011-29559-C02-01,2) and by Consejería de Educación de la Junta de Castilla y León (project VA322A11-2)

to give expression to the message. ToBI[2] is at present the most widely used prosodic annotation system. It distinguishes different pitch accents in terms of their linguistic function (phonological aspect) and their characteristic shape, either as monotonal accents, namely H(igh) and L(ow), or as a combination of these two (phonetic aspect).

Automatic ToBI accent type identification has been recently investigated with the application of a novel classification technique that allowed us to achieve a 70.8% identification accuracy rate in multi-class scenarios [5]. Binary decisions regarding presence versus absence of pitch accent reach 90% [1,9]. One of the reasons why it is difficult to overcome this recognition rate is the high level of uncertainty concerning the labelers' judgements of some ToBI annotations. The study reported in [8] supports this fact empirically, in which different transcribers are asked about the pairs of labels they find most confusing. Furthermore, some ToBI labeled corpora (the Boston University Radio News Corpus [10] among them) include notes of the transcribers stating that a second label could also be used for tagging a given accent. Taking such ambiguity into consideration, in this paper we present an experiment in which we adapted the classifier presented in [5] so that more than one accent type is assigned to each word. Various alternative pitch accents are offered to reproduce the uncertainty exhibited in the labelers' judgements.

Manual prosodic labeling is a costly task, requiring substantial time by the transcriber. Well-trained human labelers are needed to perform an activity whose duration has been estimated to take from 100-200 times real time [13]. Offering the human labeler automatic predictions for them to correct or validate is a useful strategy that has been successfully tested in previous work [13]. In the present paper, we show that assisting human transcribers with multiple ToBI prosodic labels assigned by our classifier allows a significant reduction in manual transcription time.

In the case of large speech corpora several labelers work in parallel on different parts of the corpus in order to be more efficient. If we also want the process to be effective, in a context where there is a high level of uncertainty in the labelers' judgements, we must ensure that they all follow the same labeling criteria. In this paper, we demonstrate that assisting human transcribers with ToBI prosodic labels predicted by our classifier also implies an improvement in consistency among them.

The structure of the paper is as follows. First, we present the experimental procedure (section 2); we then provide a detailed description of the corpus used (section 2.1), the automatic classifier (section 2.2), the labeler team and the labeling procedure (section 2.3), and the metrics used to assess the procedure (section 2.4). The results reveal both a significant reduction in manual transcription time and an improvement in consistency among transcribers (section 3). The conclusions are presented in section 4.

2 Experimental Procedure

2.1 Corpus and Parameterization

We used the Boston University Radio News Corpus [10]. This corpus includes labels separating phonemes, syllables and words. Accents are marked with a ToBI label and a position. We take into account the 7 more frequent types of pitch accent tones: H*, L+H*, !H*, H+!H*, L+!H*, L*, and L*+H, discarding other undetermined marks like * or *?. Inspired in previous works [1,14] we aligned the accent tones with respect to the word. We used all the utterances in the corpus with TOBI labels, from all the speakers (f1a, f2b, f3a, m1b, m2b and m3b). The total number of samples that have been considered are: H*: 7587, L+H*: 2383, !H*: 2144, H+!H*: 586, L+!H*: 638, L*: 517, L*+H: 44 and no-accent: 13868.

Similar features to other experiments reported in the state of the art have been used [1]. They concern to frequency, energy, duration and pseudo-grammatical information POS. Furthermore, the impact of alternative prosodic features that represent the evolution of the F0 contour has been considered. We included a set of coefficients representing the fitting Bézier function that stylizes the F0 contour and Tilt parameters (see [5] for details).

2.2 Automatic Labeling

The complex multi-class classification problem is divided into several simpler ones by means of pairwise coupling. We propose to combine two-class classifiers to achieve the multi-class classification, because two-class problems provide high accuracy results. Besides, the complementarity of Artificial Neural Networks (ANN) and Decision Trees (DT) classifiers has been exploited to improve the final system, combining their outputs using fusion methods (see figure 1).

The pairwise coupled approach divides a given multiclass classification problem into binary classification sub-problems whose results must be combined to get the final classification result [7,15]. According to this approach, by $\hat{P}(l|x,\lambda^{k}_{l,m})$ we mean an estimation of the probability $P(y = l|x, \quad y = l \vee m)$, where l and m are two different prosodic labels; x is the input of the classifier (in our case the prosodic features described in the previous section); y is the class label;$\lambda^{k}_{l,m}$ is a pairwise classifier of the type k (in our case a decision tree or a neural network) trained to separate the classes l and m.

From these estimators we build $\hat{P}(l|x,\lambda^{k})$, obtained with the classifier of type k, by using fusion operation:

$$\hat{P}(l|x,\lambda^{k}) = \bigotimes_{\substack{l,m=1..C \\ l \neq m}} \hat{P}(l|x,\lambda^{k}_{l,m}) \tag{1}$$

where C is the number of classes or prosodic labels and $\bigotimes$ is the fusion operator.

We step forward to fuse the results of K independent type of classifiers so that the final estimation of $P(l|x)$ would be $\hat{P}(l|x)$ computed as:

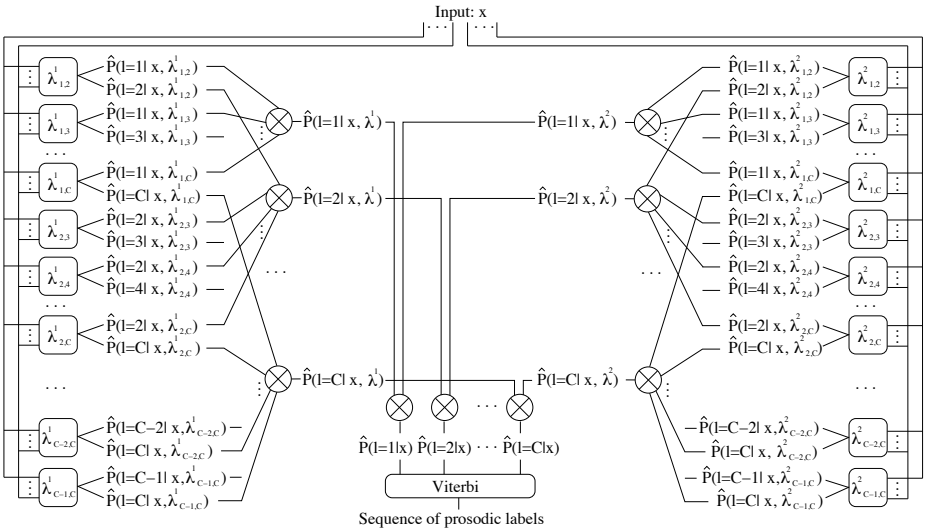


Fig. 1. Diagram of the classification procedure used in experiments.

$$\hat{P}(l|x) = \bigotimes_{k=1..K} \hat{P}(l|x,\lambda^k) \tag{2}$$

The system proposed can be seen graphically in Fig. 1. There are as many classifiers as combinations of pairs of C classes: $\frac{C\cdot(C-1)}{2}$. Each classifier, $\lambda^k_{l,m}$, provides the a posteriori probabilities estimation $\hat{P}(l|x,\lambda^k_{l,m})$ and $\hat{P}(m|x,\lambda^k_{l,m})$. The results of the classifiers are fused as described in (1) and (2). Finally, the classification rule selects the label l^* so that $l^* = \text{argmax}_l\, \hat{P}(l|x)$

In [5] we used Neural Networks as λ^1 and Decision Trees as λ^2. The fusion of results is based on the product rule and the best sequence selection is done by applying the Viterbi algorithm. This proposal, together with an adequate feature extraction, which includes the use of Tilt and Bézier parameters, allows us to achieve a total classification accuracy of 70.8% for pitch accents, 84.2% for boundary tones and 74.6% for break indices on the Boston University Radio News Corpus.

The output of the classifier is the label that obtains the highest decision score, and a number of other potential labels that are typically confused with the one selected are discarded. In this work we decided to offer the evaluation team the first three labels which were ranked as most likely for them to decide on the appropriate one. Table 1 shows how the degree of accuracy of the system increases when more than one label is taken into account. As the accuracy increases from 70.8% to 92.4%, the degree of confidence of the labelers is also expected to increase as well.

Table 1. Accuracy of pitch accent tone classification using 1 label, 2 labels and 3 labels

	1 label	2 labels	3 labels
H*	72.5%	92.1%	97.2%
L+H*	25.3%	73.4%	87.7%
!H*	35.2%	63.9%	74.9%
H+!H*	12.1%	34.6%	49.5%
L+!H*	6.0%	27.3%	45.6%
L*	11.4%	44.1%	55.9%
L*+H	0.0%	0.0%	4.5%
none	91.0%	96.5%	98.9%
Total	70.8%	86.8%	92.4%

2.3 Labeling Team and Procedure

The three transcribers who participate in this study have extensive experience with the ToBI labeling system. Indeed, one of them (T1 from now) has been involved in professional labeling projects in which she was requested to label more than one hour of a corpus for Text-To-Speech purposes.

The transcribers were asked to perform the tagging task in two different scenarios: with and without automatic prosodic labeling. Praat[4] software was used in both scenarios. In the assisted scenario, the manual labeler was confronted with TextGrid files containing five tiers: one with the orthographic transcription, three with different ToBI labels, and one tier which was empty. Transcribers had previously been informed that the labels in the tiers were ranked, being the most probable according to the automatic classifier the one in the top tier. Transcribers had to fill in the bottom tier with a number indicating which of the above tiers contained the most appropriate label (see Figure 2). If none of them seemed adequate, the transcribers supplied their own label. In the unassisted scenario, only two tiers were provided: one with the orthographic transcription and one empty tier to be filled in by the labelers. Table 3 summarizes the quantity of items that have been labeled in both scenarios.

2.4 Metrics

In order to measure the consistency of the judgments of the different labelers, we used both the *pairwise transcriber agreement* and the *kappa index*. We briefly describe these metrics in this section (see [12] for a more formal and detailed description).

The *pairwise transcriber agreement* was measured by counting the number of labeling agreements for all pairs of transcribers. That is, 4 transcribers (T1, T2, T3, T4) would produce 6 possible transcriber pairs (T1T2, T1T3, T1T4, T2T3, T2T4, T3T4). The criterion is conservative: if 3 out of 4 transcribers agree, only

[4] http://www.fon.hum.uva.nl/praat

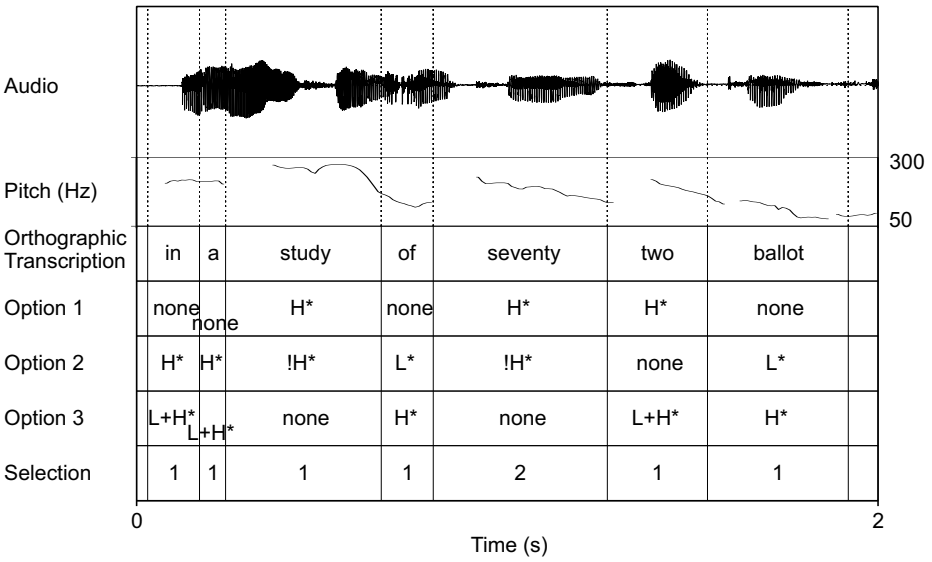


Fig. 2. Praat interface in the assisted labeling scenario.

3 out of 6 pairs will match, making the agreement rate 50% (agreement = agree / (disagree + agree)).

The Cohen's kappa, which works for two raters, and Fleiss' kappa[4], an adaptation that works for any fixed number of raters, improve upon the pairwise transcriber agreement in that they take into account the amount of agreement that could be expected to occur by chance. If a fixed number of people assign numerical ratings to a number of items then the kappa will give a measure for how consistent the ratings are. The kappa, κ, can be defined as, $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$. The factor $1-\bar{P}_e$ gives the degree of agreement that is attainable above chance, and, $\bar{P}-\bar{P}_e$ gives the degree of agreement actually achieved above chance.

3 Results and Discussion

Table 3 compares the degree of inter-transcriber consistency in both the assisted and the unassisted scenario with results from other consistency tests found in the state of the art. The global consistency rate among transcribers increases from 0.51/63.9% in the unassisted scenario to 0.55/67.0% in the assisted one. Table 2 shows that consistency increases in each pair, reaching more than 5% in the pair T1-T3.

Table 4 shows the use made by the transcribers of the different options. As can be seen, they mainly select the label corresponding to the top tier, namely, the prediction ranked first by the classifier. There are differences among the transcribers: whereas T2 and T3 use the option *Other* more frequently than T1, the latter resorts to the first option more often than T2 and T3, 71% vs. 57% and

Table 2. Inter-labeler agreement expressed as kappa index/pairwise inter-transcriber agreement. T1, T2 and T3 are the transcribers.

	T1-T2	T1-T3	T2-T3
Un-assisted	0.44/60.3%	0.46/62.6%	0.59/68.7%
Assisted	0.48/62.9%	0.54/67.8%	0.60/70.2%

Table 3. Global inter-transcriber agreement results contrasted with results reported by other studies. The numbers in the column **Pitch Accents** are the κ index and the pairwise inter-transcriber rate (as a percentage). **T** is the number of labelers, **W** is the size of the corpus in words and **S** is the number of speaking styles. *(na)* means the information is not available. The last rows of the table have been extracted from [3].

CORPUS	**T**	**W**	**S**	**Pitch Accents**
This work unassisted	3	299	1	0.51/63.9%
This work assisted	3	383	1	0.55/67.0%
Cat-ToBI [3]	10	264	4	0.462/61.17%
Am_ToBI(fe)[12]	4	644	2	0.69 / 71%
Am_ToBI(ma)[12]	4	644	2	0.67 / 72%
E_ToBI[11]	26	489	4	na / 68%
E_ToBI[16]	2	1594	1	0.51 / 86.57%
G_ToBI[6]	13	733	5	na / 71%

Table 4. Transcribers' use of the different pitch accent options expressed as a percentage, for the assisted scenario.

	First	Sec.	Third	Other	Doubt	Empty
T1	71%	20%	7%	1%	0.4%	0.0%
T2	57%	27%	3%	10%	0.4%	3.6%
T3	67%	18%	4%	11%	1.2%	2.8%

67% respectively. As for the option *Doubt*, the transcribers barely use it, which reflects self-confidence in their judgements. Finally, the label *Empty* corresponds to words with more than one stress.

Table 5 contains the inter-transcriber agreement with respect to the original labeling of the Boston Corpus. T1 has the highest agreement rate, which evidences that she is not only well-trained but also more experienced than the other two labelers.

Table 5. Consistency with the original labeling of the Boston Corpus expressed as kappa index/pairwise inter-transcriber agreement. T1, T2 and T3 represent the transcribers, and BC is the original transcriber of the Boston Corpus.

	T1-BC	T2-BC	T3-BC
Un-assisted	0.62/74.8%	0.50/63.4%	0.53/66.3%
Assisted	0.57/70.9%	0.50/63.6%	0.52/66.2%

Table 6. Consistency with the automatic labeling expressed as kappa index/pairwise inter-transcriber agreement. T1, T2 and T3 correspond to the transcribers. BC is the original transcriber of the Boston Corpus. AS is the automatic system classifier.

	BC-AS	T1-AS	T2-AS	T3-AS
Un-assisted	0.56/71.8%	0.55/71.8%	0.40/57.5%	0.44/61.9%
Assisted	0.48/66.5%	0.57/72.4%	0.41/58.4%	0.52/67.8%

The results presented in Tables 2, 3 and 5 demonstrate that computer assisted prosodic labeling introduces a bias into the labeling process of the human transcriber. Table 5 shows that the presence of automatic labels has an effect on the human experts: T1 reduces her agreement rate with respect to the original labeling. As can be observed in tables 2 and 3, both the inter-transcriber consistency and the global consistency increase because the labelers are likely to be influenced by automatic tagging.

In order to compute the savings in labeling time when the assisted scenario is used, we computed the ratio between labeling time and real time of the utterances that have been processed. The results obtained are not identical for the three human transcribers: the labeler T1, with the highest level of expertise, did not improve her labeling time, the mean value of the ratio being 55.3 if assisted by the classifier and 55.4 in the unassisted scenario. Greater time savings are observed, however, for the other labelers: the ratio is 55.95 in the assisted scenario and 67.5 when they were not computer assisted (these results show statistically significant differences when the Student's t-test is used: p-value = 0.026). These figures mean that one of these labelers could end her task about three days

before when tagging an one hour duration corpus working four hours per day (more than four hours tagging ToBI is not recommended).

Table 6 illustrates the consistency of the automatic labeling compared to the manual labelers' judgements: the value AS (automatic system) represents the first option of the three pitch accents proposed in the assisted scenario. The automatic predictions had an agreement rate as accurate as manual labelers with regard to the original tagging of the Boston Corpus (BC). In fact, only T1 has higher rates: 74.8% vs. 71.8% in the unassisted scenario and 70.9% vs. 66.5% in the assisted one. Taking into account that automatic labels can be enriched either with a degree of certainty of the prediction or with other alternative labels, we can conclude that the technique used in the automatic classifier mirrors the behaviour of the human transcriber, whose tagging, far from being utterly reliable, often results in inter-transcriber disagreement.

4 Conclusions

In this paper we have presented a computer-assisted technique for prosodic labeling of speech corpora. An automatic classifier provides the human transcriber the most likely accent types for each word, and the transcriber chooses or validates the most appropriate label.

The system has proved to be efficient, with a reduction of about 17% in manual transcription time. The effectiveness of this automatic classifier is also evident, since it reaches a success rate with respect to the original labeling which is similar to, or even higher than, those obtained by the human transcribers. Moreover, global inter-transcriber consistency improves when automatic predictions offered by the classifier are used as reference.

References

1. Ananthakrishnan, S., Narayanan, S.: Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. IEEE Transactions on Audio, Speech, and Language Processing 16(1), 216–228 (January 2008)
2. Beckman, M., Hirschberg, J., Shattuck-Hufnagel, S.: The original ToBI system and the evolution of the ToBI framework. In: Jun, S.A. (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 9–54. Oxford University Press, New York (2005)
3. Escudero, D., Aguilar, L., Vanrell, M., Prieto, P.: Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. Speech Communication (54), 566–582 (2012)
4. Fleiss, J.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
5. González-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., Cardeñoso-Payo, V.: Improving Automatic Classification of Prosodic Events by Pairwise Coupling. IEEE Transactions on Audio, Speech and Language Processing 20(7), 2045–2058 (September 2012)

6. Grice, M., Reyelt, M., Benzmüeller, R., Mayer, J., Batliner, A.: Consistency in Transcription and Labelling of German Intonation with GToBI. In: Proceedings ICSLP. pp. 1716–1719 (1996)
7. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The annals of Statistics 26(2), 451–471 (April 1998)
8. Herman, R., McGory, J.: The conceptual similarity of intonational tones and its effects on intertranscriber reliability. Language and Speech 45, 1–36 (2002)
9. Ni, C.J., Liu, W., Xu, B.: Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features. In: Proceedings of Interspeech 2011. pp. 2017–2020 (2011)
10. Ostendorf, M., Price, P., Shattuck, S.: The Boston University Radio News Corpus. Tech. rep., Boston University (1995)
11. Pitrelli, J.F., Beckman, M.E., Hirschberg, J.: Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proceedings of ICSLP. pp. 123–126 (1994)
12. Syrdal, A., McGory, J.: Inter-transcriber reliability of ToBI prosodic labeling. In: International Conference on Spoken Language Processing (ICSLP). pp. 235–238 (2000)
13. Syrdal, A.K., Hirschberg, J., McGory, J., Beckman, M.: Automatic ToBI prediction and alignment to speed manual labeling of prosody. Speech Communication 33, 135–151 (2001)
14. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2(4), 469–481 (October 1994)
15. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (December 2004)
16. Yoon, T., Chavarría, S., Cole, J., Hasegawa-Johnson, M.: Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In: Proceedings of Interspeech, Jeju. pp. 2729–2732 (2004)

Language Identification based on a Discriminative Text Categorization Technique

Miguel A. Caraballo, Luis F. D'Haro, Ricardo Cordoba, Rubén San-Segundo, José M. Pardo

Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid - E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040-Madrid, Spain

{macaraballo,lfdharo,cordoba,lapiz,pardo}@die.upm.es

Abstract. In this paper, we describe new results and improvements to a language identification (LID) system based on PPRLM previously introduced in [1] and [2]. In this case, we use as parallel phone recognizers the ones provided by the Brno University of Technology for Czech, Hungarian, and Russian languages, and instead of using traditional n-gram language models we use a language model that is created using a ranking with the most frequent and discriminative n-grams. In this language model approach, the distance between the ranking for the input sentence and the ranking for each language is computed, based on the difference in relative positions for each n-gram. This approach is able to model reliably longer span information than in traditional language models obtaining more reliable estimations. We also describe the modifications that we have being introducing along the time to the original ranking technique, e.g., different discriminative formulas to establish the ranking, variations of the template size, the suppression of repeated consecutive phones, and a new clustering technique for the ranking scores. Results show that this technique provides a 12.9% relative improvement over PPRLM. Finally, we also describe results where the traditional PPRLM and our ranking technique are combined.

Keywords: Language Identification, n-gram frequency ranking, discriminative rankings, text categorization, PPRLM

1 Introduction

Currently, one of the most used technique in Language identification (LID) is the phone-based approach, like Parallel phone recognition followed by language modeling (PPRLM) [3]. In PPRLM, the language is classified based on statistical characteristics extracted from the sequence of recognized allophones.

In spite of the high LID accuracy results obtained by PPRLM, the accuracy is reduced because PPRLM does not model correctly long-span dependencies (i.e. to use high order n-gram language models) probably due to an unreliable estimation of the n-gram probabilities. We propose to use a ranking of occurrences of each n-gram with higher n-grams, in a similar way to [4] and [5] where the ranking is applied to written

text. Although the information source is very similar to PPRLM (frequency of occurrence of n-grams), results are much better, as we will see.

This paper is a continuation of the work done in [1] and [2] but tested on a new database with 4 languages and with new improvements to the ranking algorithm. Section 2 describes the system setup and basic techniques. In Section 3, the basic n-gram ranking technique and the new discriminative n-gram ranking are described, together with the results considering all the new alternatives considered. Finally, conclusions and future works are presented in Section 4.

2 System description

2.1 Database

We have used the C-ORAL-ROM database [6], which consists of spontaneous speech for 4 main Romance Languages: Spanish, French, Portuguese, and Italian. This database is made up of 772 spoken texts with more than 120 hours of speech and around 300K words for each language.

The database transcriptions and annotations were validated by both external and internal reviewers. The database includes recordings in two different types: formal and informal. The formal recordings consist of three different contexts: natural (e.g. political speech, teaching, preaching, etc.), media (e.g. talk shows, news, scientific press, etc), and telephone (e.g. private and human-machine). The informal recordings include monologues, dialogues, and conversations in familiar and public contexts.

We needed to do several tasks to adapt the database to our experiments and recognition system: a) Most of the sound files were sampled to 22,050 Hz @ 16 bits and some others to 11 KHz @ 16 bits, all of them were sub-sampled to 8 KHz @ 16 bits in order to use them with the acoustic models of our recognizer. b) Some recordings in the database were too long (i.e. longer than 10 minutes) so they were split into shorter files. We also eliminated sections with noises, c) finally, we generated random recording lists in order to avoid any kind of bias at training. Table 1 shows the number of sentences in the database that we have finally used. The average sentence length is 6.2 seconds.

Table 1. Number of sentences by language

	Spanish	**French**	**Italian**	**Portuguese**
Sentences	17634	16474	19074	17946

2.2 General conditions of the experiments

In our previous work, we used two phoneme recognizers, for Spanish and English, with context-independent continuous HMM models. Now, we present the results using three phone recognizers in Czech, Hungarian and Russian developed by Brno

University of Technology, which are based on using neural network classifiers and were trained on the SpeechDat-E databases.

The phoneme recognizers output contains many relevant errors for several reasons: a) there is a mismatch between the recognizers' languages and the four languages to be identified; b) the recordings still contain different kind of noises, background music, etc., and very spontaneous speech; c) the acoustic models were not adapted to this database. So, there is a clear mismatch in the languages and in the channel conditions. At least, improvements obtained with our techniques will be more evident, as we will see.

In order to increase the reliability of the results presented in the next sections, we performed a cross-fold validation, dividing all the available material in 9 subsets: 5 subsets to estimate the LMs, 2 subsets to estimate the Gaussian classifier, 1 subset for development, and 1 subset for test.

2.3 Description of PPRLM

Nowadays, PPRLM is one of the two typical approaches to language identification. The main objective of PPRLM is to model the frequency of occurrence of different allophone sequences in each language. The technique can be divided into two stages. First, several parallel phone recognizers take the speech utterance and outputs a sequence of allophones corresponding to the phone sets used for each recognizer. Second, the sequence of allophones is used as input to a bank of n-gram language models (LM) in order to capture phonotactics information. The LM module provides the probability that the sequence of allophones corresponds to a given language.

The main advantages of PPRLM are: a) since it uses many recognizers, it is possible to cover most of the phonetic realizations of every language. b) It is possible to have phone recognizers modeled for languages different to the languages that we have to identify, which is especially useful in situations where the training data is not enough to obtain reliable models. On the other hand, PPRLM presents a major weakness: the data sparsity limits the LMs ability to model long span information.

For score normalization, given the good results obtained in [7], we decided to continue using a Gaussian Classifier as a backend. These classifiers also benefit from applying normalization of the scores (e.g., the T-norm normalization). In our system, we use what we call "differential scores", which applies a similar normalization.

Regarding solutions for the problem of including long span information to the language models, in [8] they describe slight improvements on the LID rate when using the skip-gram technique. In [5] they present LID experiments on written text for six languages using three different kinds of LM: Markov models, trigram frequency vectors, and n-gram text categorization, with good results for the last technique. Finally, in [9] an interesting algorithm for selecting high order n-grams based on dynamically keeping the most frequent ones is presented but the selection is not based on discriminative information among languages. In our case, we have used and extended the n-gram text categorization technique proposed in [4].

2.4 PPRLM Results for LID

One problem with the PPRLM approach is that it is affected with a bias that appears in the log-likelihood score for the languages considered. To tackle this issue, we proposed in [7] to use a Gaussian classifier instead of the usual decision formula applied in PPRLM. With all the scores provided by every LM in the PPRLM module we prepare a score vector. With all the sentences in the training database, we estimate a multi-Gaussian distribution for each language. In recognition, the distance between the input vector of LM scores and the Gaussian distributions for every language is computed, using a diagonal covariance matrix, and the distribution which is closer to the input vector is the one selected as identified language.

One important conclusion of our work in [7] is that, instead of absolute values, we need to use differential scores: the difference between the score obtained by one LM and the average score obtained by the other 'competing' languages: (SCi' = SCi – Aver(SCj, j≠i)). We applied it to unigram, bigram and trigram separately, with 4 languages x 3 phone decoders x 3 n-grams = 36 features in total in the feature vector.

The average result in LID for PPRLM is 29.89% error rate. It seems a high rate, but, as we mentioned in Sections 2.1 and 2.2, the performance of the acoustic models is really poor and the sentences average length is short.

3 N-Gram Frequency Ranking

In this section, we will describe the original text categorization technique and the modifications that we have made to improve it, as well as the algorithm for selecting the most discriminative n-grams and to rank them.

3.1 Description of the Basic Technique

In [4], an interesting technique that combines local information (n-grams) and long-span information (collected counts from the whole utterance) is described. In summary, during training, the original technique proposes the creation of a ranked template with the N (typically 400) most frequent n-grams (up to n-grams of order five) of the character sequences in the train corpus for each language sorted by occurrence.

During the evaluation, a dynamic ranked template is created for the phoneme sequence of the recognized sentence following the same procedure. Then a distance measure (OOP, Out-Of-Place) is applied between the input sentence template and each language dependent template previously trained. The distance for a given ranking T is calculated using Equation 1.

$$d^T = \frac{1}{L}\sum_{i=1}^{L} abs(pos\ w_i - pos\ w_i^T) \tag{1}$$

Where L is the number of n-grams generated for a given input sentence. If an n-gram does not appear in the global ranking (meaning that it is not in the top n-grams selected) it is assigned a maximum distance, i.e., the size of the ranking. The selected language is the one that presents the lower distance between templates. Fig. 1 shows an example of one of the templates created in our system using the English phoneme set and the template created for the unknown sentence.

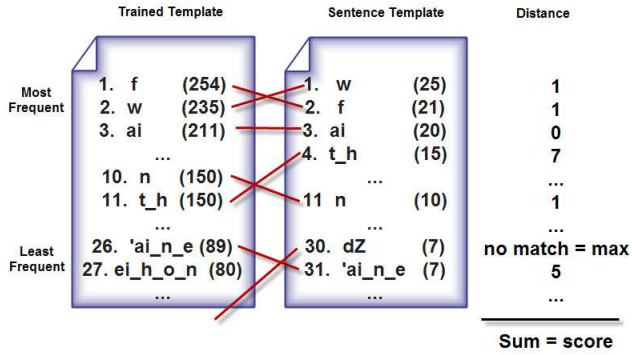


Fig. 1. Example and calculation of distance score using a ranking of n-grams as proposed by [4]

3.2 Our baseline for N-Gram Ranking

In [2] we described several modifications that we made on the basic technique proposed in [4]. We will mention here just the most relevant one. We applied what we called the "golf score". As the number of occurrences of the n-grams in the input sentence is very low, most n-grams have the same number of occurrences and should have the same position in the ranking. It is the same as a ranking in golf (the sport): all players with the same number of strokes share the same position. Fig. 2 shows an example of the modification applied to the original template, which provided a 2.4% relative improvement.

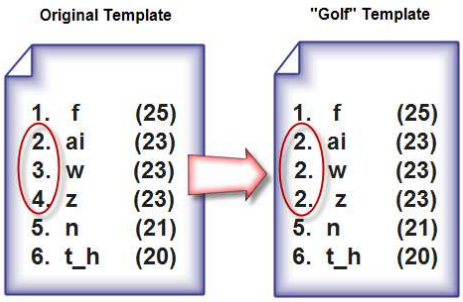


Fig. 2. Ranking template modification with "golf score"

3.3 N-Gram Discriminative Ranking

Inspired in the work presented in [10], where better LID results could be obtained using the most discriminative units, we thought that we should introduce the same concept in the ranking creation process; therefore, we decided to give more relevance (higher positions) in the ranking to the items that are actually more specific to the language that is being identified, i.e. n-grams with a high frequency in one language but with zero or low frequency in the competing languages.

In our work we propose a variation of tf-idf. After the original global rankings are created, we have the number of occurrences of each n-gram: $n_1(w)$ = occurrences of n-gram w in the current language, and $n_2(w)$ = the occurrences of w in the competing language, where T is the whole set of ranking templates created for each language.

$$N_1 = \sum_{\forall w \in T_i} n_1(w) \qquad N_2 = \frac{1}{|T-1|} \sum_{\forall w \in T_j, j \neq i} n_2(w) \tag{2}$$

As the number of total occurrences will be different for each language and n-gram order, before the subtraction a normalization is needed to have comparable amounts. Being N_1 the sum of all occurrences for the current language and N_2 the average for the competing languages (see Equation 2):

$$n_1'(w) = \frac{n_1(w) \times N_2}{N_1 + N_2} \tag{3}$$

$$n_2'(w) = \frac{n_2(w) \times N_1}{N_1 + N_2} \tag{4}$$

Another important issue is to use a threshold value for these normalized values (Equations 3 and 4), i.e., to discard the n-grams that were below a threshold as non-representative. In our case, we obtained an optimum using 6-6-2-2-2 (from 1-gram to 5-gram from left to right). Then, we considered several alternative formulas (shown in Table 2) with the same philosophy as tf-idf for the final number of occurrences used to assign the final position in the ranking (which we will call n_1'').

Table 2. Different formulas for discriminative selection used in the experiments

1	$n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')$
2	$n_1'' = \log(n_1') * (n_1' - n_2') / (n_1' + n_2')$
3	$n_1'' = (n_1' - n_2') / (n_1' + n_2')$
4	If $n_1' > n_2' \rightarrow n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')^2$ Else $n_1'' = n_2' * (n_1' - n_2') / (n_1' + n_2')^2$
5	If $n_1' > n_2' \rightarrow n_1'' = (n_1' - n_2') / (n_1' + n_2') * [1+\log(n_1' / (n_1' + n_2'))]$ Else $n_1'' = (n_1' - n_2') / (n_1' + n_2') * [1+\log(n_2' / (n_1' + n_2'))]$
6	If $n_1' > n_2' \rightarrow n_1'' = (n_1' - n_2') / (n_1' + n_2') * \text{sqrt}(n_1' / (n_1' + n_2'))$ Else $n_1'' = (n_1' - n_2') / (n_1' + n_2') * \text{sqrt}(n_2' / (n_1' + n_2'))$

In our experiments, we obtained similar results from formulas 4, 5, and 6. In all cases, they improved the baseline with a 5% reduction in LID error rate. We decided to use formula 4 as it meant a more consistent improvement considering several experiments, also because it was slightly better for 3-grams and 4-grams which are the most discriminative ones.

All the formulas that we propose (3-6) normalize the values between 1 and -1: where 1 means that the n-gram appears in the current language but not in the other competing ones (n_2'=0), therefore indicating that the n-gram is especially relevant for that language; -1 means just the opposite (n_1'=0).

3.4 Suppression of repeated consecutive phonemes

In the phone recognizer outputs, several consecutive phonemes are the same, especially for silences. These repeated silences affect the n-gram calculation, especially for 4-gram and 5-gram. Therefore, we decided to suppress them, leaving one instance of each phoneme.

Table 3. LID error rate and improvement obtained after removing repeated consecutive phonemes.

	Original output	Filtered output	Relative improvement
1-gram	46.89	45.35	3.30%
2-gram	35.71	32.07	10.18%
3-gram	30.65	27.56	10.11%
4-gram	33.12	30.03	9.31%
5-gram	46.83	44.07	5.89%
All	30.31	26.91	11.19%

We can see that this decision provides a significant improvement, so we will use it on all experiments.

3.5 Influence of the template size

In our first experiments we worked with template sizes up to 4000. One thing that we observed with this database is that sizes could be increased more drastically with success. Another obvious point is that the template size should be different depending on the n-gram order, as the number of units is clearly different. Therefore, we run a series of experiments varying the template size for the different n-gram orders, which we can see in Fig. 3.

We can draw several conclusions from the figure: 1-gram and 2-gram are not affected by these template sizes, which could be expected as the number of items is usually below 3,000. The saturation points are: 14,000 units (3-gram), 34,000 (4-gram), 66,000 (5-gram). This could also be expected as there are more input n-grams

as the order increases. The best result now is 26.50% LID Error rate, slightly better than the 26.91% that we obtained with the previous non-optimized template sizes.

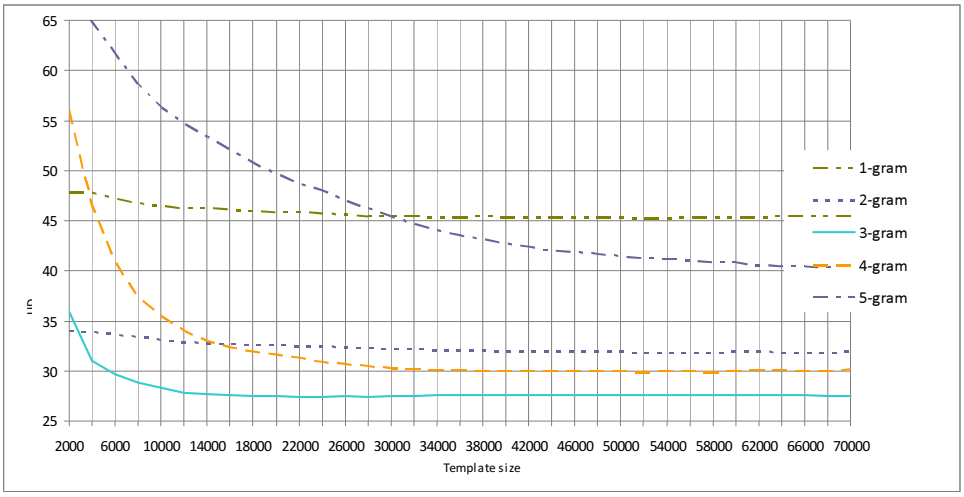


Fig. 3. Template size results for each n-gram

3.6 Clustering of ranking scores

In the previous section, we observed that results tend to saturate as the template size increases. We think this is due to the following: if the size is too big, the out-of-rank units (the ones that do not appear in training or have low values in the scores used for creating the ranking, the n_1'' from Section 3.3) receive a great penalty in the distance formula: they are assigned the last position in the template (the template size).

So, we propose to use another approach. With the n_1'' scores (normalized between 1 and -1) we made a clustering so that units with similar scores share the same position in the ranking. The clusters are created using the classical k-means algorithm. This way we can handle a larger number of units, but still apply a reasonable penalty to unseen units: now, they are assigned the total number of clusters, which is quite a smaller value.

In Table 4 we can see the results, including the total number of clusters obtained with k-means, the total number of units in those clusters and the LID error rates. The results values correspond to the average of all k-fold experiments (hence, the decimal values). We can see that the approach is worse for 1-gram and 2-gram, which is normal as we reduce the number of units, and hence, the precision. But it is extremely nice that the performance increased for 4-gram and slightly for 5-gram. We also observe that the final number of clusters is small for 4-gram and especially 5-gram, which is the result of having many units with very similar scores.

Table 4. Clustering of ranking scores.

	No. units	No. clusters	LID error rate with clustering	LID error rate without clustering	Improvement
1-gram	51.2	27.8	63.41	45.48	-38.71%
2-gram	1,510.2	733.7	31.81	31.96	0.98%
3-gram	20,015.6	1,985.4	27.63	27.34	-0.60%
4-gram	55,397.4	1,225.9	**28.96**	29.85	**2.08%**
5-gram	40,875.5	392.8	37.60	40.38	0.10%
All			26.23	26.50	-0.33%

The conclusion is that we should use our regular templates for 1-gram and 2-gram, and the clustering approach for the rest.

3.7 Combination of PPRLM and N-Gram Discriminative Ranking

We checked whether the two techniques were complementary, so we fused them. The summary of results is as follows:

- PPRLM: 29.89%
- N-gram Discriminative Ranking: 26.23% (12.2% relative improvement)
- PPRLM + N-gram: 26.04% (12.9% relative improvement)

So, the fusion of both of them does not provide significant improvements in these experiments.

4 Conclusions and Future Work

We have shown that the n-gram Frequency Ranking approach overcomes PPRLM due to the longer span that can be modeled, especially for the effect of the 4-gram, and partially of the 5-gram information. The following issues have been crucial:

- n-gram discriminative rankings with the normalized value for the number of occurrences are able to overcome PPRLM (12.9% relative improvement).
- Using a ranking score normalized between 1 and -1 provides significant improvements.
- The ranking size should be different for n-gram orders.
- The clustering of ranking scores provides further improvements as it lets to consider all units appearing in training that are above a threshold.
- The suppression of repeated consecutive phonemes provides a significant improvement, 11.19%.

As future work, we will work with NIST LRE databases, and we will fuse the results with systems based on acoustic information.

5 Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02) and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects.

6 References

1. Cordoba, R., D'Haro, L.F., et al. "n-gram Frequency Ranking with additional sources of information in a multiple-Gaussian classifier for Language Identification". V Jornadas de Tecnología del Habla, pp. 49-52, 2008. Bilbao, Spain.
2. Cordoba, R., D'Haro, L.F., et al. "Language Identification based on n-gram Frequency Ranking". Interspeech 2007, pp. 354- 357.
3. Zissman, M.A., "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech&Audio Proc., v. 4, pp. 31-44, 1996.
4. Cavnar, W. B. and Trenkle, J. M., "N-Gram-Based Text Categorization", Proc. 3rd Symposium on Document Analysis & Information Retrieval, pp. 161-175, 1994.
5. Vatanen, t., Väyrynen, J. and Virpioja, S. "Language Identification of Short Text Segments with N-gram Models". Int. Conf. on. Language Resources and Evaluation (LREC'10), 2010.
6. Cresti, E. et al. "The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages". IV Int. Conf. on Language Resources and Evaluation, 2004.
7. Córdoba, R., et al. "Integration of acoustic information and PPRLM scores in a multiple-Gaussian classifier for Language Identification". IEEE Odyssey 2006.
8. Navratil, and J. Zühlke, W. "Double bigram-decoding in phonotactic language identification". ICASSP, Vol. 2, pp. 1115–1118. 1997.
9. Penagarikano, M. et al. "A dynamic approach to the selection of high order n-grams in phonotactic language recognition".Acoustics, Speech and Signal Processing (ICASSP), 2011.
10. Nagarajan, T., and Murthy, H. A. "Language Identification Using Parallel Syllable-Like Unit Recognition". ICASSP, pp. I-401-404. 2004.

BINARY POSITION ASSIGNMENT OF TWO KNOWN SIMULTANEOUS ACOUSTIC SOURCES

Rupayan Chakraborty, Climent Nadeu, and Taras Butko

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
Campus Nord, C/ Jordi Girona, 1-3, Barcelona, Spain
{rupayan.chakraborty, climent.nadeu, taras.butko}@upc.edu

Abstract. When two acoustic sources are simultaneously active in a meeting-room scenario, and both the position of the sources and the identity of the two sound classes are estimated, still there is the problem of assigning each source position to one of the two sound classes. We found this problem in the real-time system implemented in our smart-room, where one of the sounds, when time overlapping exists, is always speech, and the source positions are relatively well separated in space. The position assignment system is based on partial source separation achieved using null steering beamforming with either two or three microphones, followed by a log-likelihood-ratio-based classifier. Both frequency dependent and frequency invariant beamforming are explored in this paper. The experimental results from offline simulations show a high assignment rate for most of the acoustic events overlapped with speech.

Keywords: Source position assignment, null steering beamforming, acoustic event detection, simultaneous sounds.

1 Introduction

Since the human activity is reflected in a rich variety of acoustic events, the automatic detection and localization of sounds can help to describe the social activity that takes place in a meeting-room environment [1] [2]. Recently, a real-time system was deployed in our smart-room to detect and localize acoustic events (AEs) that may occur simultaneously with speech [3]. That temporal overlap of sounds poses a challenge to the acoustic event detection (AED) and acoustic source localization (ASL) systems, especially because only signals captured by a set of distant microphones are used.

The ASL system, which employs a set of 24 microphones regularly distributed in the room, was designed to estimate up to two source positions [3]. The AED system, which uses the signal from one microphone, deals with the overlapping problem at the level of models [1], i.e. an additional acoustic model is considered for each AE overlapped with speech, so the number of models is doubled. In this way, the AED system can detect the presence of either one event or two events overlapped in time [3].

Once we have the output of the AED system, there is the need of a posterior identification of which one of the source positions given by the ASL system corresponds to speech and which one corresponds to the detected AE. That requires an additional position assignment module, which is developed and tested in this paper.

2 Acoustic scenario and database

We have aimed to design the system to operate in the general case i.e. speech ("sp") source is produced at one (unknown) position, say P_1, and the acoustic event source is produced at the other (unknown) position, say P_2, inside the room . The main motivation of the work is to resolve the identification ambiguity of the overlapped sounds. In the working scenario, it is assumed that speech is always produced at one side of the room (left or right), and the other AE are produced at the other side.

For training, development and testing of the system, we have used, as in [3], the audio part of a publicly available multimodal database produced in our lab. There are 8 sessions of audio data for isolated acoustic events. Each session was recorded with all the 6 T-shaped microphone arrays each one having 4 microphones. The total number of acoustic event classes is 12, including speech [3]. The overlapped signals of the database were generated adding those AE signals recorded in the room with the speech signal, also recorded in the room, both from all the 24 microphones. To do that, for each AE instance, a segment with the same length was extracted from the speech signal starting from a random position, and added to the AE signal. The mean power of speech was made equivalent to the mean power of the overlapping AE.

3 Basic scheme of the position assignment system

The block diagram of the whole detection and localization system is presented in Fig. 1. The two-source AED block and the two-source ASL block are those described in [3]. The AED system uses the model based approach, which showed best results [4], and provides either one or two detected events. The ASL system, which employs all 24 microphones, is based on a SRP-PHAT localization method, and also provides either one or two source positions. The position assignment block is detailed in Fig.2. In the most general case, we have two detected events, i.e. "sp" and E (one of the 11 possible AE), and two source positions: P_1 and P_2. The position assignment system assigns each one of the two detected events to each one of the two positions. It consists of two null steering beamformers (NSB), which work in parallel, plus a subsequent classifier. Each NSB should be designed to steer to one of the two positions provided by the ASL system while placing a null in the direction of the other position. Thus, the contribution of one of the two simultaneous sounds to the NSB output is expected to be higher than its contribution to the signal input. There are two kinds of inputs to the beamformers: 1) multi-microphone signals, and 2) two sets of position coordinates. The classifier takes a binary decision based on a log-likelihood ratio, as it will be explained in Section 5.

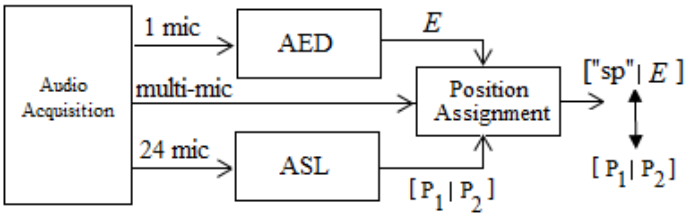


Fig. 1. Block diagram of the whole system

4 Null steering beamformers

Null steering beamforming (NSB) is one of the earliest, but potentially very useful beamforming techniques. It belongs to a class of very popular and widely used beamforming technique called multiple side lobe cancellers (MSC) [5] [6].NSB adapts the sensor array pattern by steering the main beam towards the desired source and placing nulls in the direction of the interference sources [7]. Here, beamforming is based on the prior knowledge of the direction of the desired and interference sources. In our work, two types of beamformers are explored: 1) frequency dependent beamformer (FDB), and 2) frequency invariant beamformer (FIB). Both of them have their advantages and disadvantages. The main advantage of FDB is that it has less computational load. It assumes the input signal is a sinusoid (narrow band), so beamforming is dependent on its frequency. As the audio signals are broadband, the selection of a proper frequency value becomes necessary. This is the main drawback of FDB. On the other hand, for FIB, the beam patterns do not vary with frequency what makes this approach robust, although the frequency invariance adds complexity to the algorithm. The details of these two types of beamforming are discussed in the following subsections.

4.1 Frequency dependent beamformer

In our scenario we have only two sources; one is the target and the other is an interference that has to be nulled. So a first-order NSB is needed which uses multiple microphones [9]. The output of a beamformer can be expressed for each time instant as

$$\mathbf{x}^{\mathrm{H}}\mathbf{W} \tag{1}$$

where, **X** is the Nx1 (N microphones) input signal vector, and **W** is the Nx1 weight vector, which is the solution of the equation (2)

$$\mathbf{S}^{\mathrm{H}}\mathbf{W} = \mathbf{e}^{\mathrm{H}} \tag{2}$$

where, **e**=[1 0] and **S**=[$s_{m,n}$] is a 2xN matrix whose elements are defined by

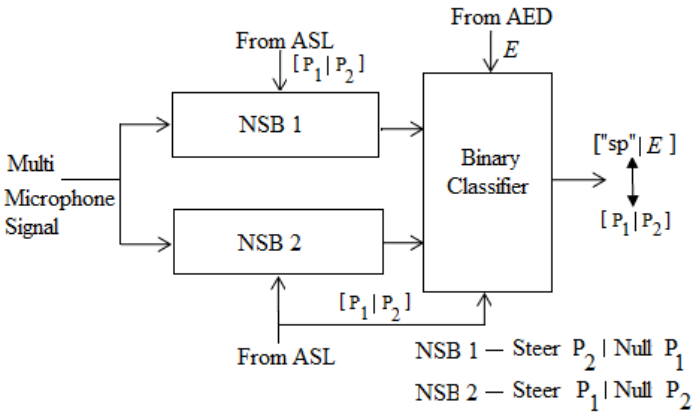


Figure 2. Position assignment system

$$s_{m,n} = \exp\left(j2\pi f (n-1)\left(d \cos\theta_m\right) / v_s \right) \tag{3}$$

where n is the microphone index; $s_{1,n}$ is the steering vector for the target source, and $s_{2,n}$ is the steering vector for the interfering source (which are a function of the directions of arrival (DOA) θ_m), f is the operating frequency (narrow-band beamforming is assumed), d is the spacing between the two microphones, and v_s is the velocity of sound in m/sec.

4.2 Frequency invariant beamforming

The main objective of this method is to introduce a broadband beamforming by decoupling the spatial selectivity from the frequency selectivity through a parameterization of the filter coefficients. Once the decoupling is done, the frequency invariant response is obtained by choosing the same coefficients for multiple frequencies. Generally, the broadband frequency invariant beamforming mostly requires a continuous aperture of sensors with a large number of microphones on a specific geometry. But this method uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones. Moreover, it is possible to steer the resulting frequency invariant response by combining it with the spherical decomposition of the beam pattern [8]. The filter array response in the frequency domain is represented as

$$h(f,\theta) = \sum_{n=1}^{N} c_n(f) g_n(f,\theta) \tag{4}$$

where N denotes the number of real sensors in the array geometry. This equation can be seen as a parameterization of the filter array response for each frequency f with

coefficients $c_n(f)$ and basis function $g_n(f,\theta)$. A modification of the coefficients $c_n(f)$ will affect the frequency and spatial response simultaneously because $g_n(f,\theta)$ depends on both frequency f and direction of arrival θ. The goal of this frequency invariant beamforming is to find a new parameterization for $c_n(f)$,

$$c_n(f) = \sum_{l=1}^{L} b_{nl}(f)\,\hat{c}_l(f) \tag{5}$$

such that $b_{nl}(f)$ is the basis transform that converts the array response into frequency invariant array response by replacing N sensors indexed by n by a new set of L virtual sensors indexed by l.

$$\sum_{n=1}^{L} g_n(f,\theta) b_{nl}(f) = \tilde{g}_l(\theta) \tag{6}$$

These virtual sensors are frequency invariant. Note that the coefficients $\hat{c}(f)$ are the same for all frequencies. For the calculation of $b_{nl}(f)$ in equation (5) it is required to use an elegant analytic inversion formula which imposes some limitations: $l<N$ and all the sensors are placed equidistant in a defined geometry. To overcome the problem of the restricted number of sensors and the restrictions on sensor locations imposed by the analytic inversion, a least squares solution is used. The arrival angle in this case is discretized by θ_q, $q=1, 2,....,Q$ and the equation (6) can be rewritten in matrix form as

$$\mathbf{G}(f)\mathbf{B}(f) = \tilde{\mathbf{G}} \tag{7}$$

where $[\mathbf{G}(f)]_{qn} = g_n(f,\theta_q)$, $[\mathbf{B}(f)]_{nl} = b_{nl}(f)$ and $[\tilde{\mathbf{G}}]_{ql} = \tilde{g}_l(\theta_q)$. The least squares solution to this equation is defined by

$$\mathbf{B}(f) = \mathbf{G}^{\dagger}(f)\,\tilde{\mathbf{G}} \tag{8}$$

where $\mathbf{G}^{\dagger}(f)$ is the pseudo inverse of $\mathbf{G}(f)$. After calculation of this basis transform $\mathbf{B}(f)$, it is made steerable by spherical harmonics. To present the whole methodology in a compact form, we can write the array response in the following way,

$$h(f,\theta) = \mathbf{g}(f,\theta)\mathbf{B}(f)\mathbf{D}(\alpha,\beta,\gamma)\,\hat{c}(f) \tag{9}$$

where $\mathbf{g}(f,\theta)$ is a vector of N sensor responses to a plane wave, and $\mathbf{B}(f)$ uncouples the spatial response from the spectral response. $\mathbf{D}(\alpha,\beta,\gamma)$ steers the spatial response into an arbitrary direction and it is the rotation matrix commonly known as Wigner rotation matrix which is a function of azimuth α, elevation β and the spin angle about z-axis γ. Therefore, the basis transform has a rotation to any directions specified by (α,β,γ) and $\hat{c}(f)$ defines it's spatial profile for each frequency separately. Choosing the same coefficients for all frequencies yields an array response that is frequency invariant.

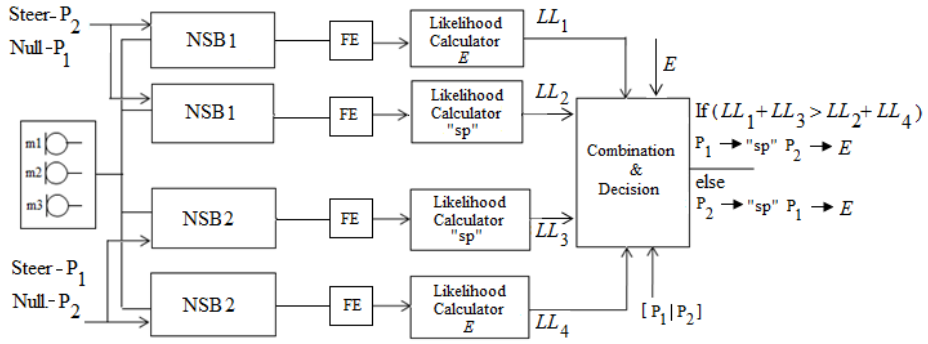


Figure 3. Complete scheme of the binary position assignment system

5 Position assignment system

The complete scheme of the position assignment (PA) system used in this work is shown in Fig. 3. As described in Section 3 and Fig. 2, the PA system consists of beamforming based partial signal separation, followed by likelihood ratio based binary classification. In order to get the most from the available information, the classifier includes both a speech model and an AE model. Indeed, we can work separately with only either the speech model based classifier or the AE model based classifier. All the options are tested here.

As shown in Fig. 3, two different NSBs are needed, in general, for each parallel path of the system, one adapted to the AE and the other to speech. Moreover, when FDB is used, a different beamformer is needed for each acoustic event class, since the frequency *f* is tuned to a particular event.

The classification part of the PA system consists of a feature extraction (FE) block at the first stage, a likelihood calculator at the second stage, and a binary decision block at the last stage [4]. Features are extracted from the audio signals with frame length of 30ms, and 20ms frame shift. In this work, we have used frequency-filtered log filterbank energies (FF-LFBE) which are applied in speech recognition [10]. In our experiment, we considered 32 dimension feature vectors (16 FF-LFBE and its first temporal derivatives). There is a set of two likelihood calculators for each parallel path, one of them to calculate the log likelihood for the AE label (*E*) provided by the AED system and the other to calculate it for speech. As in [3], here we employ Hidden Markov Model based likelihood calculators, where Gaussian Mixture Models (GMM) are used to calculate the emission probabilities [11]. 64 Gaussian components with diagonal covariance matrix are used per model. 11 HMM are trained with the isolated events using the Baum-Welch algorithm [11]. The HTK toolkit is used for training and testing this HMM-GMM system [12].

Finally, the decision block first combines and then compares the log-likelihood values from all the calculators. When both the speech and the AE model based classi-

fiers are used, the decision is taken based on the combination of the log-likelihood values from the output of all four classifiers, as depicted in Fig. 3. The classifier's outputs are combined in the decision block using the product rule (addition of log likelihoods).

If only either the AE or the speech based classifier is used, the path with the largest log-likelihood output is indicating the decision that has to be taken. Let's illustrate it with a particular case. Assume P_1 truly corresponds to speech and P_2 to the one of the acoustic event E. When the classifier is trained only with isolated acoustic events, it is expected to get comparatively higher log-likelihood from the output of NSB1 (LL_1) than from the output of NSB2 (LL_4). In the case when the classifier is trained with clean speech, we expect to get comparatively higher log-likelihood from the output of NSB2 (LL_3) than from the output of NSB1 (LL_2). If that is the case, the decision for both types of classifiers is taken that speech is at P_1 and E is at P_2, which is the correct decision.

5.1 Evaluation metrics

To design and evaluate the performance of the system, we define the Position Assignment Rate (PAR) metric for a given AE as the quotient between the number of correct decisions and the total number of occurrences in the testing database. We have also considered a second metric called Diff_LL, which is defined as the difference between the log-likelihood values computed by the calculators for correct assignment. While maximization of the PAR is our main criterion for evaluation, Diff_LL has been used with a second level priority. When tuning the frequency f, sometimes occurs that the PAR for an AE is the same for several frequency values, since the number of AE occurrences is not high enough. Then, the frequency f that maximizes Diff_LL is chosen. In fact, that difference can be considered as an estimate of the degree of source separation carried out by the beamforming system when a correct assignment is made.

5.2 System implementation

In this paper, we consider two types of beamforming at the front end of the PA system: FDB and FIB. Their design requires the DOA angles corresponding to the target and the null, i.e. the DOA from source positions P_1 and P_2. For both of them, we have tried with approximate angles from visual inspection during recording. Regarding the AE position, we have also used the output of a one-source ASL system.

The formulation of the FDB approach is comparatively simpler than the FBI one, but, as it assumes a narrow-band signal, its design requires frequency tuning to get optimal results. So the beam patterns for this type of NSB are heavily dependent on both the DOA and the frequency f. As we are looking to design our system for dif-

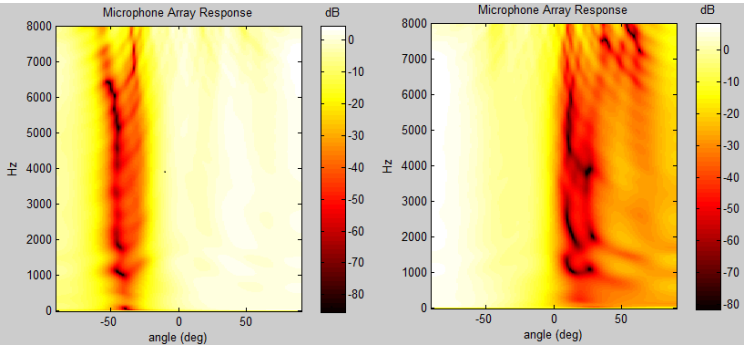


Figure 4. Beam patterns of frequency invariant beamforming. Left: null towards speech. Right: null towards AE

ferent types of acoustic events with diverse spectra, the choice of a frequency value for each specific event becomes necessary. Here, we adopted an exhaustive search technique for that. We varied the frequency from 100 Hz to 8 KHz with intervals of 100 Hz, and observed the performance of the system for each acoustic event separately.

6 Results and discussion

In our offline experiments, we use data from seven sessions (S02-S08) for training the system and tuning the frequency *f* in the FDB case, and one session (S01) is left for testing. Six of the seven training sessions are used for training the AE models, and the remaining session is used for testing each frequency value. Cross-validation is used to have a better statistical meaning. The frequency that shows the best average behavior according to the PAR metric is chosen.

In all the experiments reported in this paper, instead of using the AED system output to set the AE model used by the likelihood calculators in the classifier, the ground truth was used, i.e. the errors from the AED system are not affecting the measure of accuracy of the position assignment system in this preliminary work.

Let us recall that the FDB needs a separate beamformer with an optimized frequency for each acoustic event. But for FIB, a single beamformer is sufficient for all acoustic events. In our experiments, we have observed that, for the FDB case, the use of the ASL output leads to higher PAR scores than using the AE source positions specified during recording. This may be due to the fact that those values are coarse approximations. In fact, there is also some variation in the position of the different instances of a given AE class. For the FIB case, the opposite behavior has been observed. The reason for that may be the broad main lobe of the beam pattern (as can be seen in Fig. 4). Additionally, in the FIB case, the NSB is not specific of each event class since that broad beam encompasses all the angles from where AE were produced

in the recordings sessions. Consequently, this simplifies the overall system, as it does not require any external ASL block.

To assess the system performance, several experiments are conducted. The testing results are obtained with all the 8 sessions (S01-S08) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all other 7 sessions are used for training. Results obtained using the microphone array T-6, for both metrics, averaging over all the testing dataset, are presented in Table 1. A comparison is done between FDB and FIB, with three types of configurations: 1) classifier is based only on AE models, 2) classifier is based only on speech model, and 3) a likelihood combination of the AE model based and the speech model based classifiers. Note that, for both types of beamforming, the likelihood combination improves the overall performance of the system.

From Table 1, we observe that FDB works better than FIB in terms of the PAR metric, though the relative difference is small in the AE case. The best results is obtained from the frequency dependent beamformer with likelihood combination. In that case, the performance of the FIB based system is higher in terms of Diff_LL, which indicates that it achieves better signal separation when the assignment is correct. The slight advantage offered in our experiments by the FDB system in terms of PAR has to be balanced by the fact that it requires a different beamformer for each AE class, and needs frequency tuning for each of them. Additionally, it requires a more precise source position.

Table 1. Performance scores of the position assignment system

	FDB			FIB		
	AE based	Speech based	LL comb.	AE based	Speech based	LL comb.
PAR (%)	87	88.3	89.4	86.2	81	87.1
Diff_LL	1.38	1.23	1.48	1.44	1.07	2.52

7 Conclusion

An attempt to resolve the source identification ambiguity that appears when an acoustic event overlapped with speech is detected is carried out in this paper. A position assignment system has been proposed and tested, which consists of a beamforming based partial signal separation technique, followed by a likelihood ratio based binary classifier and a decision block. Two types of beamforming techniques are compared in this paper. The frequency dependent beamformer gives slightly better results than the frequency invariant one, but it is more dependent on the particular conditions of the production of acoustic events in the room. Both acoustic event models and speech models are used for the likelihood computation, and a product rule based combination of their likelihood calculators improves the performance of the system.

8 Acknowledgement

This work has been supported by the Spanish Government funded project SARAI, TEC2010-21040-C02-01.

References

1. Temko, A., Nadeu, C.: Acoustic event detection in meeting-room environments. Pattern Recognition Letters.vol. 30/14, pp 1281-1288, Elsevier. (2009)
2. Zhuang, X., Zhou, X., Hasegawa-Johnson, M., Huang, T.S.: Real-world acoustic event detection. Pattern Recognition Letters, vol. 31/12, pp. 1543-1551, Elsevier. (2010)
3. Butko, T., Gonzalez Pla, F., Segura, C., Nadeu, C., Hernando, J.: Two-source acoustic event detection and localization: online implementation in a smart-room. Proc. EUSIPCO. (2011)
4. Chakraborty, R., Nadeu, C., Butko, T.: Detection and positioning of overlapped sounds in a room environments. Proc. Interspeech. (2012)
5. Van Veen, B. D., Buckley, K. M.: Beamforming: A Versatile Approach to Spatial Filtering. ASSP Magazine, IEEE. Volume: 5 Issue: 2, pp. 4-24, April. (1988)
6. Applebaum, S. P.: Adaptive Arrays. IEEE Trans. Antennas and Propagation, 24:585-595, September.(1976)
7. Hoshuyama, O., Sugiyama,A.: Robust Adaptive Beamforming.in Microphone Arrays: Signal Processing Techniques and Applications.Ed. Brandstein, M., Ward, D.: New York: Springer. (2001)
8. Parra, L. C.: Steerable Frequency-Invariant Beamforming for Arbitrary Arrays. Journal of the Acoustical Society of America. 119 (6), pp. 3839-3847. June. (2006)
9. Teutsch, H., Elko,G. W.: First and second order adaptive differential microphone arrays. Proc. IWAENC. (2001)
10. Nadeu, C., Macho, D., Hernando, J.: Frequency &timefiltering of filter-bank energies for robust HMM speech recognition. Speech Communication, vol. 34, pp. 93-114. (2001)
11. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice Hall.(1993)
12. Young, S. et al.: The HTK Book (for HTK Version 3.2), Cambridge University. (2002)

Emotions recognition using binary fingerprints

Xavier Anguera, Esperança Movellan* and Miquel Ferrarons*

Telefonica Research,
Edificio Telefonica-Diagonal 00, 08019, Barcelona, Spain,
xanguera@tid.es

Abstract. Recognizing emotions in speech recordings is a topic rapidly gaining popularity in the scientific community which has interesting challenges in terms of research and potential applications in a real world scenario. In this paper we focus on the identification of the predominant emotion in a speech excerpt among a set of possible (previously known) emotions. In most prior art research this problem is approached by using acoustic features directly derived from the audio (sometimes expanded to several thousand components) and using standard machine learning techniques to construct emotion models for each of the known emotions. In this paper we propose the use of an alternative approach recently introduced for speaker modeling which brings the feature space, modeling and classification to the binary space. Experiments show that the proposed method obtains very competitive results using standard (low dimensional) features.

Keywords: Emotions recognition, binary fingerprints, classification, modeling

1 Introduction

Recognition of the emotions reflected in an audio signal is a quite recent field of research that is quickly gaining popularity. Indeed, this topic is not only interesting from the point of view of research, with several novel unanswered challenges, but also in a real world scenario, with multiple applications in call-centers or monitoring user interaction with automated dialog systems.

Most research performed until now in this area [5, 6] has been addressed as a classification problem where an input utterance is classified among a set of two (neutral vs emotional) or several known emotions (e.g. anger, sadness, happiness, etc.). Statistical machine learning techniques used to model each of the emotions range from Support Vector Machines (SVM) [7], Neural Networks (NN) [3], Gaussian Mixture Models (GMM) [1], Hidden Markov Models (HMM) [2], among others. Currently, in order to obtain state-of-the-art results usually

*Esperança Movellan and Miquel Ferrarons were visiting students from Universitat Pompeu Fabra at the time of this work.

hundreds (or even thousands) of features are stacked together into feature vectors that capture the dynamics and acoustic characteristics of each emotion. Many times a single feature vector ends up being used for an input utterance as a result of the application of various statistics on short-term features. Such high number of features is usually not easy to interpret and it is not ensured whether all information contained in the short-term feature vectors is retained or not. In order to reduce features dimensionality some works like [4] perform feature selection to select those features that are most prominent in differentiating among emotions, which still does not answer concerns about interpretability and retainment of information.

In this paper we propose a departure from common practices in two ways. On the one hand, we use a totally different modeling approach by constructing binary fingerprints from each emotion and then comparing them entirely in the binary domain. This modeling approach was initially been proposed in [10] as a way to model speakers by their voices and is the first time it has been successfully applied to emotion recognition. On the other hand, our system uses standard acoustic features (MFCC 39-dimensional at this point, but this is not a limitation of the system) which are not compressed into a single feature vector at feature level. Even though such features are of much lower dimensionality and only represent a very local portion of the signal, with binarization we are able to indirectly incorporate long term information and successfully model emotions. Furthermore, given that the emotion fingerprints are represented by a binary vector it is much simpler to interpret how an emotion is represented and see the differences between different emotions.

Experimental results show that our approach can achieve very competitive results in comparison to results obtained with a well known emotion recognition system available online, while using much simpler features.

2 Binary Emotion Classification Algorithm

The proposed emotion classification algorithm is based on the binary speaker modeling algorithm initially proposed in [10]. It is comprised of three main modules as seen in figure 1. First, an offline module trains a KBM (binary Key Background Model) on as much acoustic data as possible. The KBM model is a GMM model whose function is to perform the change of base between the input acoustic features into binary form. Second, emotion models are trained given the data available for each emotion. To do so, the data is first converted into binary form using the KBM model and then a binary fingerprint is obtained. The resulting binary fingerprint models are small to store and, as will be seen later, fast to compare. Third, an online test module uses the KBM model to transform any input utterance into binary domain and then compares them to the emotion fingerprint models, selecting the one with highest similarity. Further details in each of these modules is given next.

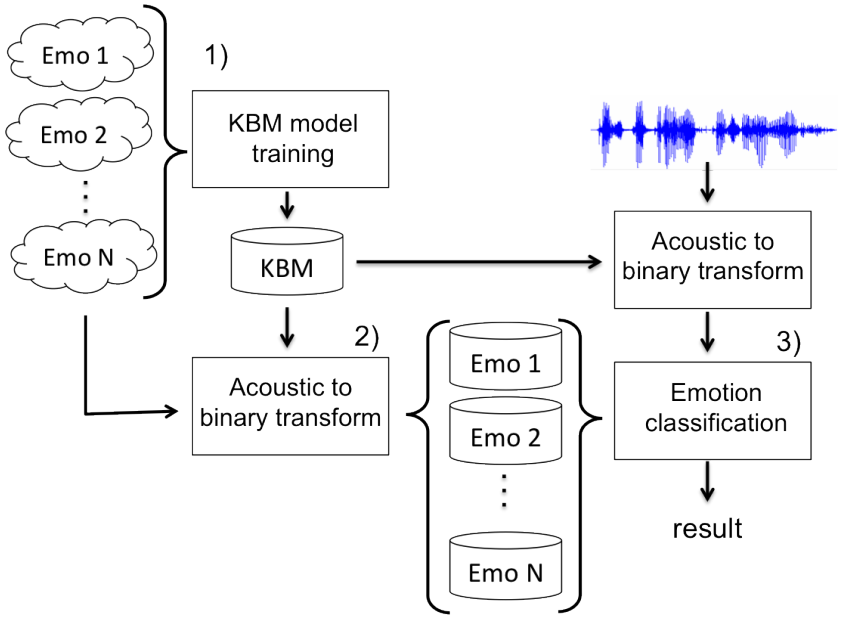


Fig. 1. *Binary emotion classification algorithm, comprised of three main modules: 1) KBM training; 2) emotion fingerprint training; 3) classification of input utterances.*

2.1 KBM Training

The KBM model is used to transform acoustic feature vectors into binary form. The KBM model differs from a standard GMM model in the objective the model is created for. While in a generative model when training one wants to optimize the representation of the data given the model, the objective of the KBM model is to optimally segment the acoustic space so that small differences in the input feature vector derive in very different binary fingerprints only if these differences are important. Many alternatives can be used to train the KBM model (for example [11]). In this paper we use a simple EM-ML training using Gaussian initialization through Gaussian splitting. In order to train the KBM appropriately we need training data that covers all emotions we want to recognize, all within similar acoustic conditions that the ones we test on. In this paper we do not use any prior information regarding which emotion each training file belongs to and use all acoustic data in a single training pool. For this reason we considered using the same database used in test, as there is not danger of data overfitting.

Once we have the trained KBM model, for each acoustic feature vector (usually extracted every 10ms from the audio signal) we use the model to obtain a binary feature vector whose dimensionality corresponds to the number of Gaussians used in the KBM model, existing a direct relation between each Gaussian in the model and a position in the feature vector. We set to 1 the positions in the vector corresponding to the closest Gaussians to that feature vector, where

distance to a Gaussian is measured in terms of likelihood of feature vector x being modeled by that Gaussian λ, i.e. $Lkld(x|\lambda)$. Intuitively, by doing this binarization we are reducing the information we keep from the feature vector to encode only the region in the acoustic space where such feature occurs. Such regions come from quantizing the acoustic space using the KBM model so that when we encounter a high density of feature vectors in training data we quantize the space more densely, and vice-versa. After some experimentation we got to the conclusion that regardless of any other parameters or application, selecting around 5 closest Gaussians per feature vector always achieved always good performance. This base transformation results in a very sparse binary vector that can be easily compressed if need be for storage or transmission.

2.2 Emotion Binary Fingerprints Computation

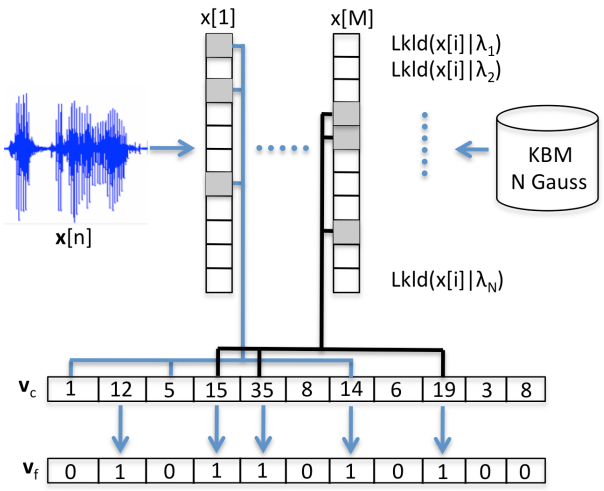


Fig. 2. *Fingerprint training algorithm.*

Figure 2 shows the overall process used to compute a binary fingerprint for an emotion given some acoustic training data. This is the same process followed to convert an input utterance into a single vector binary form for classification into an emotion. In fact, the block in Figure 1 titled *acoustic to binary transformation* corresponds to the algorithm is Fig. 2 preceded by an acoustic feature extraction step to convert the input audio signal into the desired intermediate acoustic feature vectors.

The different steps involved in the algorithm shown in Fig. 2 are explained next. First, given a set of training utterances summing M acoustic feature vectors $x[i]|i = 1 \ldots M$ we convert them into binary form using the KBM model as explained in the previous section. We then count how many 1 values we obtained

for each position in the M binary vectors and save such count into the v_c vector, which has the same dimension as the binary feature vectors (i.e. the number of Gaussians in our KBM model). The v_c vector contains the counts histogram indicating which Gaussians are closest to the M given input feature vectors. The last step in the process is to convert these counts vector into a binary form. In this case we choose the 25%-highest count dimensions to become 1, setting to 0 the rest. The resulting binary fingerprint has the size of the number of Gaussians in the KBM and is usually several orders of magnitude smaller than any traditional GMM or SVM model.

Note that what this step actually does is an averaging of the information obtained at feature level regarding where in the acoustic space the input features mostly occur. This is very similar to the averaging performed by state-of-the-art emotion recognition systems at feature level by representing the whole utterance using several functionals computed from the set of acoustic features extracted from the signal, like in our case, at short-term intervals. The main difference here is that as each bit in the fingerprint directly represents the activation status of a given KBM Gaussian, and each Gaussian represents a regions in the multidimensional space where the input acoustic feature vectors reside, we can directly obtain a relationship of which acoustic regions are most prevalent for each one of the emotions, being able to more easily interpret the results.

2.3 Emotions Identification

Given an input signal we want to classify into the emotion its most closely reflects, we first convert such signal into binary form and then compare it with the set of finite emotions for which we have binary fingerprints. As mentioned before, the binarization of the test utterance is done in the same way as with the training of emotion fingerprints, i.e. first obtaining an intermediate acoustic feature vector representation (the same in which the KBM model was trained) and then mapping these features into binary space and summarizing the test utterance using a single binary vector.

The comparison of the input binary vector with the available emotions is done in the binary space by using the similarity measure shown in 1. Similarities obtained range from 0 to 1, the higher the closest. Note that given that all the processing after binarizing the feature vectors is performed on the binary space, it is very fast to compute.

$$S(\mathbf{v}_{f1}, \mathbf{v}_{f2}) = \frac{\sum_{i=1}^{N}(v_{f1}[i] \wedge v_{f2}[i])}{\sum_{i=1}^{N}(v_{f1}[i] \vee v_{f2}[i])} \quad (1)$$

3 Experiments

In this section we describe the experimental setup and the results of our proposed system, compared to state-of-the-art alternatives.

3.1 Experimental Setup

In order to test the proposed algorithm we used the publicly available emoDB database [8]. The database is freely available online [1] and consists of 535 short recordings where seven emotions (Anger, Boredom, Disgust, Fear, Happiness, Sadness, and Neutral) are simulated by several German actors. Given the small size of the database most researchers (including us) perform a 10-fold cross-validation experiment, where 9/10 of the database is used for training models and the rest is used for testing, which results in 10 tests whose results are averaged to obtain a single overall result. For each partition we use an equal number of files from each emotion. Note that the exact way in which partitions should be constructed is not commonly defined, therefore it is difficult to compare results among research papers as each particular distribution might give very different results. Such results are evaluated both using a non-weighted accuracy metric (i.e. the percentage of files for which the emotion has been correctly classified) and a weighted accuracy metric where each individual emotion accuracy is inversely weighted by its a priori probability within the database. When using emoDB the weighted metric is necessary as a different number of files is available for training each emotion.

To validate our proposed system we compared the results with the open-source OpenEar package [7], which has been proven to obtain state-of-the-art results using SVN-based models for each emotion. The software was downloaded from [2] and default parameters were used to perform model training and recognition, with the exception of the front-end configurations that were not available in the original package, which were computed by modifying the example configuration scripts.

The selection of the acoustic features used for the experiments needs some explanation as it differs slightly between our proposed system and the openEar package. The openEar package offers several configuration files that extract a set of acoustic feature vectors ranging from PLP, MFCC, Pitch and others at fixed short-term intervals on the signal and later computes a set of functionals to obtain a single feature vector for the whole input utterance (both for training and testing utterances). The dimensionality of the resulting vectors varies according to which exact functionals are used (e.g. Extremes, Regression, Moments, Percentiles, Crossings, Peaks, Means, etc.). For these experiments we used the parameter sets labelled as *emo_base (*988 *dimensions), emo_large (*6552 *dimensions) and emo_IS09 (*384 *dimensions, used in the 2009 emotion recognition challenge [9]). In addition, we created* 2 *extra set of features to better compare results with our proposed algorithm. The first feature set corresponds to a 39-dimensional MFCC vector extracted every* 10*ms and collapsed into a single vector per utterance using the mean functional. The second feature vector corresponds to the same initial features (39-dimensional MFCC) to which we applied the same functionals used in the emo_IS09, obtaining a single vector os* 469 *dimensions.*

[1] http://http://pascal.kgw.tu-berlin.de/emodb
[2] http://sourceforge.net/projects/openart

For our system we have chosen to initially test the system using a standard feature vector used in many areas of speech processing. The feature vector corresponds to MFCC-39 which includes 12 MFCC features + energy, their deltas and double deltas. These were obtained from the signal by using a 25ms Hamming-windowed signal every 10ms. Note that these features are extracted with a different front-end than the one than comes with openEar, although all parameters have been set equally with the exception that in here we initially do not perform a mean of the input features into a single vector. In addition, in order to better compare our system with openEar, we extracted features using the openEar front-end and using the appropriate configurations to obtain the emo_base, emo_large and emo_IS09 sets, but without applying any functional and using all short-term feature vectors as input to our system. These result in feature vectors of dimension 53, 171 *and* 35 *respectively.*

3.2 Results

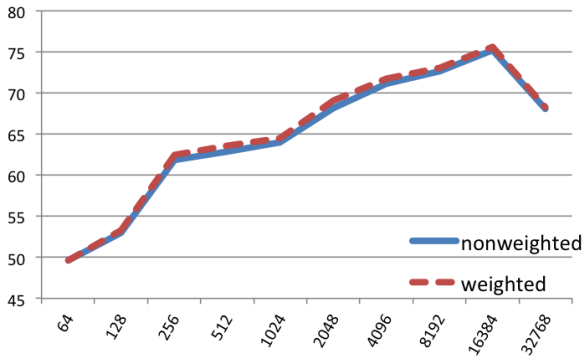


Fig. 3. *Binary fingerprint accuracy plot for different number of bits per fingerprint.*

In an initial test we evaluate the performance of our binary fingerprinting system with respect to the dimensionality of the binary vector, i.e. the size of the KBM model. To do so, we train a KBM using all data available from all partitions of the emoDB database. Then, for each partition we obtain the emotion fingerprint for all 7 *emotions and test them on the test data for that partition. Results are shown in Fig. 3 for multiple of* 2 *KBM sizes. We observe how the weighted and non-weighted accuracy metrics are both giving very similar (almost identical) results. For both cases, accuracy keeps increasing until* 2^{14}*. We hypothesize that beyond that point there might be not enough data in the training database to correctly and consistently partition the acoustic space. In fact, when performing the EM-ML iterations we observe very small Gaussian variances for many of the Gaussians, which makes them very unstable when performing the conversion to binary form. Although optimum results are obtained well beyond*

2048 *dimensions, we use this dimensionality in the comparisons from this point on as it seems to be a good tradeoff between complexity and accuracy.*

Table 1. *Comparison of results with 10-fold cross-validation testing*

system	features	weighted acc.	non-weighted acc.
OpenEar	MFCC39	56.27%	57.58%
	MFCC39 + IS_09 stats	66.52%	65.4%
	IS_09	67.77%	65.79%
	emo_base	72.45%	70.84%
	emo_large	72.24%	70.86%
Binary	MFCC39 (2048 dimensions)	68.20%	69.14%
	MFCC39 (16384 dimensions)	**75.23%**	**75.62%**
	IS_09 (2048 dimensions)	59.64%	61.63%
	emo_base (2048 dimensions)	65.20%	65.91%
	emo_large (2048 dimensions)	67.98%	66.99%

Next, Table 1, shows the comparison between openEar and our system on the different feature vectors described above. On the upper part of Table 1 we show results for openEar. As expected, results on the MFCC39 feature vector are quite poor. This might be due to the limited number of dimensions available for the SVM-based modeling and the loss of information derived from performing only an average functional on the data. When adding extremes, regression coefficients and moments (MFCC39 + the IS2009 functionals) results get much improved. Furthermore, if we use the predefined features provided by the openEar package we get the highest results, being the emo_base features the best performing with the smallest dimensionality. Note that the results for emoDB that we are obtaining are considerably lower than those advertised in papers like [7]. While we believe we are using the same setup as they do in their experiments (with the exception of the lists used for each partition) to date we have still not been able to replicate their results.

On the lower part or Table 1 we show results using our proposed binary fingerprint system on the same feature vectors. Unlike with the openEar system we are able to extract much more information out of the MFCC39 features, which obtain very good results. From all values in Figure 3 we have chosen the ones for 2048 *and* 16384 *dimensions, both obtaining results exceeding those shown for openEar. Opposite from openEar, when using any of the feature vectors that are commonly used in emotion recognition we are not able to achieve as good results as with MFCC39. We hypothesize that the reason for this is that with a reasonable binary dimensionality (*2048 *is used in these tests) it might not be enough for the KBM to properly quantize the acoustic space of such high-dimensional acoustic feature vectors, even if these might not include much usable information in some of the dimensions. It remains as future work how to be able to successfully process these high-dimensional features to take full advantage of all the information they contain.*

Over all, we can see how using simple MFCC39 features we achieve similar (or better) results that using the standard high-dimensional feature vectors currently used in emotion recognition.

4 Conclusions and Future Work

Recognizing the most prominent emotion that arises from a given acoustic utterance is a task which has recently attracted the interest of the research community and has clear implications and application in the commercial world. Typically, emotion recognizers are classifiers that select which emotion model better fits the acoustic data. Models are trained using standard machine learning approaches and feature vectors are usually obtained by stacking together several acoustically-derived parameters together, and postprocessed by functionals that model the evolution of these parameters over time. In this paper we propose a novel approach for emotion classification that has been recently introduced for speaker modeling. Its main characteristic is that emotions modeling and search for the best matching emotion is all done in the binary domain. This is the first time we demonstrate that this novel method can be suitable beyond speaker modeling and show that with simple input feature vectors we can achieve results that are in many occasions better than those of state-of-the-art emotion recognition systems.

Future work will involved testing the system with different acoustic features to find an optimal combination of features to give optimum results, and the test of the system on bigger and more challenging databases.

References

1. *I. Luengo, E. Navas and I. Hernaez: Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge. In Proc. Interspeech 2009, Brighton, UK, 2009.*
2. *R. Huang and C. Ma: Towards a speaker-independent Real-Time Affect Detection System. In Proc. International Conference on Pattern Recognition, 2006, pp.1204-1207.*
3. *M.W. Bhatti, W. Yongjin and G. Ling: A neural network approach for human emotion recognition in speech. In Proc. International symposium on Circuits and Systems, 2004, pp II-181-4 Vol. 2.*
4. *Jia Rong, Gang Li and Yi-Ping Phoebe Chen: Acoustic feature selection for automatic emotion recognition from speech. In Information Processing and Management Vol. 45, 2009, pp. 315328.*
5. *S. G. Koolagudi and K. S. Rao: Emotion recognition from speech: a review. In International Journal of Speech Technology, no. July, Jan. 2012.*
6. *M. El Ayadi, M. S. Kamel, and F. Karray: Survey on speech emotion recognition: Features, classification schemes, and databases. In Proc. International Conference on Pattern Recognition, vol. 44, no. 3, pp. 572-587, Mar. 2011.*
7. *Florian Eyben, Martin Wllmer, and Bjrn Schuller: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In Proc. Interspeech 2009, Brighton, UK.*

8. *F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss: A database of german emotional speech. In Proc. Interspeech 2005, Lisbon, Portugal, pages 1517 1520, 2005.*
9. *Bjorn Schuller, Stefan Steidl and Anton Batliner: The INTERSPEECH 2009 emotion challenge. In Proc. Interspeech 2009, Brighton, UK, 2009.*
10. *Xavier Anguera and Jean-François Bonastre: A novel speaker binary key derived from anchor models. In Proc. Interspeech 2010, Makuhari, Japan, 2010.*
11. *Xavier Anguera: Speaker Independent discriminant feature extraction for acoustic pattern-matching. In Proc. ICASSP 2012, Kyoto, Japan, 2012.*

Measuring Acoustic Reduction in Feature Space

Henrik Schulz and José A. R. Fonollosa

Departament de Teoria del Senyal i Comunicacions,
Universitat Politècnica de Catalunya
henrik.schulz@upc.edu, jose.fonollosa@upc.edu

Abstract. Modelling varying speaking style remains a challenge to state of the art speech recognition and synthesis systems. Vowel and consonant reduction have been identified as correlative to speaking style variation, but still lack a common measurement. The reduction phenomena are often observed without consideration of coarticulation and assimilation effects, and as a result of speaking rate variability. We present an analysis of acoustic reduction in Mel Frequency cepstral coefficients (MFCC) feature space of phonemes, estimate duration and determine the degree of correlation between duration reduction and feature space reduction for two different speaking styles present in broadcast news and conversational recordings. We analyse the feature space reduction of consonants and vowels in context in a syllable environment.

1 Introduction

The speaking style is an essential intrinsic variable for modelling spoken language. Its effects became more prominent in recent speech synthesis research [1], and remained challenging in recognition [2][3]. Although often categorised into planned, extemporaneous or spontaneous, it spans a large continuous scale influenced by manifold parameters. The external factors are considered situation, context, intention and listeners. These result into a composition of linguistically varying expressions, phonetically varying pronunciations and prosodically varying tone of voice. The pronunciation variation here is not meant in the strict phonological sense, but phonetically, as a varying quality, i.e. divergence of a pronunciation with minimum coarticulation and assimilation towards their maximum.

An increasing speaking rate is regarded as the fundamental cause for the deformation of the principal pronunciation. The original assumption concluded, that the higher the speaking rate, the lower the speech effort to articulate a word in its original pronunciation, consequently leading to a reduction in pronunciation, i.e. phoneme substitutions and deletions, and to a more intense coarticulation and assimilation. Effects of increased speaking rate on the acoustics of phonemes were initially studied solely on vowels analysing the trajectory of formant frequencies. The original assumption of Lindblom [4] stated that each vowel has a articulator target position. The articulator will reach the target position if there is sufficient time. However, if the duration is short, the articulators not having sufficient time (depending on the context they are embedded in)

to form the movement to the target position, that movement will be reduced. Hence, the vowels would not reach their target position but 'undershoot', which will result into different formant frequencies.

Later studies opposed above view that shorter vowel duration leads to reduction of the target position. The studies of Pols [5] showed formant trajectories that were steeper but met the same formant target values as with a reduced speaking rate. Equivalent findings were published by Moon in [6], where clear speech at larger differences in speaking rate although reducing duration did not reduce the vowel formant frequencies. Here, both studies assumed a constant speaking style. As opposed to constant speaking style, while analysing read and spontaneous speech, vowels exhibited a centralisation of formant frequencies for most vowels significantly in spontaneous speech[7].

Effects on consonants due to different speaking styles were studied also in [7] proposing several parameters: First, the difference of slope of the second formant between beginning and end of a consonant as a measure of coarticulation strength, second, the centre of gravity (COG) correlating with vocal effort, and third, energy difference between vowels and consonants which is related to intensity. Spontaneous speech exhibited significant reduction in comparison to planned speech for the first measure as well as for the second with the exception of plosives. The third measure indicated a reduction, but with a lack of statistical significance.

Word stress and contextual assimilation was shown to influence the vowel reduction in [8]. The assimilation of vowels was a more significant tendency than vowel centralisation.

Studies on different languages also resulted in opposing conclusions, e.g. in [9] a raising of vowel formant frequencies along the formant F_1 and lowering along F_2 but not a centralisation of vowels was shown for Catalan. Coarticulation effects and a certain degree of centralisation was observed for vowels in German [10].

As a remark, the mentioned phonetic vowel and consonant reduction refers to the pronunciation quality, whereas phonological reduction refers to a reduction of the number of phonemes. Although both can be related, the latter should be subject to pronunciation modelling.

All above described observations on acoustic reduction have in common a change in spectral shape. Centralisation, raising or lowering of vowel formant frequencies, assimilation and coarticulation effects, the equilibration of intervocalic energy or the change in spectral COG of consonants show a tendency of phonemes to lose their original distinct spectral characteristics the more the principal pronunciation is deformed. It remains unclear, however, to what degree neighbouring vowels and consonants (mutually influence their reduction) reduce respectively whether one exposes stronger reduction than the other. The parameters explored in the above mentioned studies are limited to a specific group of phonemes and therefore do not facilitate a continuous analysis along consecutive phonemic segments. They also disregard the full spectral length, e.g. by solely taking shifts in F_1 and F_2 into account. Apart from the assimilation ef-

fects, that were suggested in [8][10], phonetic vowel and consonant reduction has been treated as a phenomenon independent from coarticulation and assimilation effects.

As a consequence, we propose a feature space analysis (FSA) as a common measure for phonetic acoustic reduction of vowels and consonants alike. Measured in MFCC feature space, it takes the full spectral bandwidth into account, takes advantage of state of the art feature space from speech synthesis or recognition and provides a continuous scale.

In the following chapters we analyse a general degree of acoustic reduction in MFCC feature space and duration reduction across Catalan phonemes observed in three corpora of different origin and speaking style. Furthermore we evaluate contextual units such as syllables and compare the reduction phenomenon of their vowels and consonants.

2 Methods

2.1 Data

The study aims at using real life recordings from a natural environment, that are typical tasks to state of the art automatic speech recognition (ASR) systems. We employed three Catalan spoken language corpora, each of distinct origin featuring different speaking styles:

1. Catalan dictation (DI), effectively functioning as reference for the studied phenomena, and for normalisation purposes. Its utterances were read from a prepared script and therefore well planned upfront. Each sentence is lexically and grammatically complete, and without hesitations, repetitions and repairs. The speech is continuous and fluent, but not fast. The speakers are non-professional but experienced.
2. Catalan broadcast news (BN), featuring a planned and extemporaneous speaking style. The news presentations are of rather well prepared content employing professional speakers and commentators. Their phrases exhibit a rare occurrence of lexical and grammatical incompleteness, clear and distinguishable pronunciation and avoid regional accents.
3. Catalan broadcast conversations (BC), containing debates on selected topics in politics, economy or society. They feature purely conversational speech of less prepared content and presents often unprofessional but experienced public speakers. The speech features frequent hesitations, repetitions and repairs. Their phrases are often lexically and grammatically incomplete, and acoustically influenced by the course of the discussion [11].

Segments of speaker overlap, music overlap and other noises, as well as speech segments not uttered in studio environment were exempted from the original BN and BC corpora for this study.

Table 1 shows the proportions of the employed speech data for the three distinct origins. Each column denotes the number of speakers (SP), segments (SE)

Table 1: Corpora speaker distribution

	DI			BN			BC		
Gender	SP	SE	D [h]	SP	SE	D [h]	SP	SE	D [h]
male	55	7759	39:11	352	4519	10:21	441	25335	24:33
female	54	7253	36:55	172	3020	06:27	113	3848	03:51

and total duration (D) per gender and origin. Pronunciations of the orthographic transcriptions of above referred spoken language databases were derived using the SEGRE rule based grapheme-to-phoneme transcriber for Catalan [12]. The Catalan phoneme set and its SAMPA notation on which this study relies on is described in [13].

2.2 Segmentation

The phonemic analysis presupposes a segmentation of spoken utterances into its phonemic segments which were derived from automatic alignments of the utterances. Disjoint HMM tri-phone based acoustic models (AM) were iteratively estimated from each spoken language corpus using Mel frequency cepstral coefficients (MFCC) of the 10 ms shifted signal frames. The MFCC were subject to mean normalisation. The AM rely on context dependent semi-tied continuous density HMM using a 3-state topology for each tri-phone. Their emission probabilities are modelled with Gaussian mixtures sharing a common diagonal covariance matrix. A classification and regression tree ties the HMM states to generalised triphone states. Since the HMM estimation is based on the expectation-maximisation algorithm that is effectively implemented as alignment and model re-estimation, the final segmentation results from an alignment whose average likelihood is at maximum or has reached a saturation during the estimation process. Additionally to word pronunciations, syllable boundaries are provided with the grapheme-to-phoneme transcription [12]. They are assigned to the phonetic alignment and facilitate the FSA conditioned to syllable segment and type.

2.3 Feature Space Analysis and Duration

The effects of phonetic vowel and consonant reduction and the notion of increasing coarticulation and assimilation of neighbouring phonemes due to the reduction in speech effort results in increased similarity of the phonemes. This means that the distance to each other reduces respectively the distance to a common centre of the feature distribution. This is noted by the FSA that measures the size of the MFCC feature space and provides a distinct parameter for a phonemic segment in terms of its specific distance to a common mean. MFCC feature vectors are subject to cepstral mean normalisation (CMN) to reduce effects of different recording channels.

We employ speech data A that are to be analysed, and speech data N enabling a normalisation of the determined parameters of A. The speech data N are required to feature minimal deformation of the original pronunciations and need to expose a speaking style with a maximum contrast between phonemes. The analysis data A originate from the BC and BN speech corpora, while the normalisation data N originate from the DI corpus.

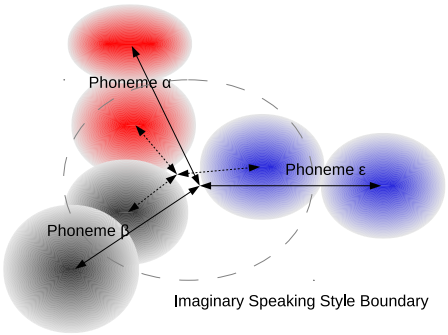


Fig. 1: Feature Space Analysis.

Figure 1 depicts the scheme of the feature spaces for two imaginary speaking styles, one exposing more reduction and similarity between phonemes in the centre than the other, both separated by a dashed line that indicates the imaginary boundary between speaking styles. The three phonemes featuring acoustic reduction are more difficult to discriminate and increase the overlap between their feature spaces while the variance remains the same. The scheme assumes a constant level of reduction for one and the same speaking style. As further contextual analysis will show, this is not necessarily the case with real speech data.

A FSA has been subject to an ASR performance study in [3] estimating the distances between the average feature vector μ_{ph} for each phoneme ph to the centre of the distribution of all phonemes for the speech data μ^A, and employs the same distance measure to the speech data N to facilitate a normalisation. The normalised FSA denotes to:

$$FSA_{ph} = \frac{||\mu^A_{ph} - \mu^A||}{||\mu^N_{ph} - \mu^N||} \tag{1}$$

As above FSA definition requires the presence of the same speaker in both data sets A, N and does not facilitate an analysis per phonemic segment (as required for a segment specific but also syllable conditioned analysis), we rephrase the estimation formula. The average feature vector of a spoken phonemic segment seg is denoted as $\mu_{ph,seg,sp}$ (sp indicates the speaker). We keep the speaker specificity while estimating the average $\mu_{ph,sp}$ of all segments of a particular

phoneme, and the resulting centre of the distribution of all phonemes μ_{sp}. The latter is defined as average $\mu_{ph,sp}$ over all phonemes ph exposed by a particular speaker sp. Instead of applying the Euclidean distance between the average over all phonemic segments of a particular phoneme ph and the centre of the distribution of all phonemes, we solely estimate a phonemic segment specific distance:

$$d_{ph,seg,sp} = ||\mu_{ph,seg,sp} - \mu_{sp}|| \tag{2}$$

We determine these distances for both analysis speech data A and normalisation speech data N, and normalise the segment specific distances of analysis speech data with the average distances per phoneme D^N_{ph}, i.e. determined from the distances $d^N_{ph,seg,sp}$ across all segments seg of the speakers sp from normalisation speech data N:

$$FSA_{ph,seg,sp} = \frac{d^A_{ph,seg,sp}}{D^N_{ph}} \tag{3}$$

Although the FSA is a language independent method, its estimates are assumed to be influenced by various language specific properties, i.e. dialect, stress, etc.

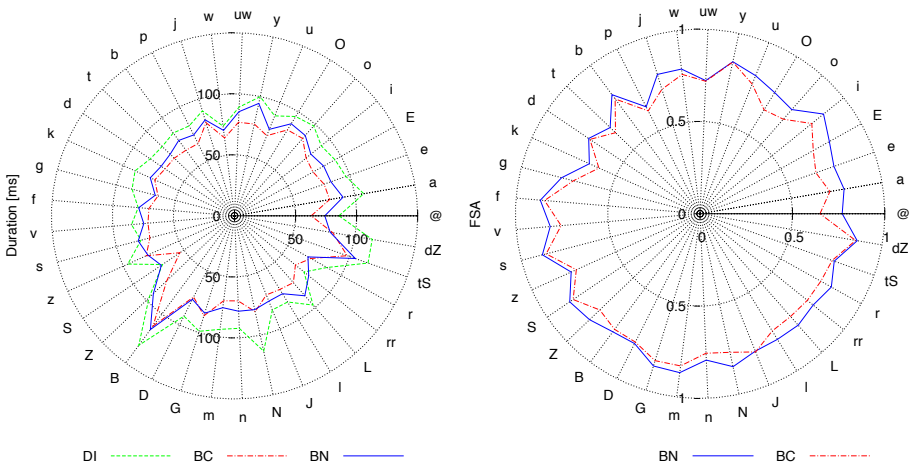
An increasing speaking rate respectively a reduction in duration is considered an essential cause for the observed acoustic reduction in vowels and consonants as outlined in chapter 1. Therefore a primary analysis of phoneme duration was carried out. It essentially relies on the number of aligned frames per phonemic segment.

3 Results : FSA on Broadcast News and Conversational Speech

For both BN and BC speech data we measured both phoneme duration and FSA for each phonemic segment, while DI speech functioned as phoneme duration reference and as FSA normalisation data.

Figure 2a displays average phoneme duration for DI, BN, and BC speech data. The duration of most phonemes of BN and BC speech reveals a statistically significant reduction in comparison with DI speech (ANOVA $p < 0.05$), while the phoneme duration of BC speech indicates the lowest duration for most phonemes. From the studies introduced in chapter 1, we conclude on a potential phonetic acoustic reduction, but also bear in mind, that even though, phonemes may reduce in duration respective are affected by higher speaking rate, the may not necessarily exhibit phonetic acoustic reduction.

Figure 2b depicts the FSA of phonemes originating from speech in BN and BC, showing a large feature space reduction in comparison to the DI planned speech reference distribution at the imaginary circle at 1. A higher FSA value implies less phonetic acoustic reduction respectively more resistance to coarticulation and assimilation. Comparing the FSA of BN and BC indicates significant



(a) Duration Analysis (b) Feature Space Analysis

Fig. 2: Phoneme duration and feature space analysis of DI, BN and BC data

differences for all phonemes (ANOVA $p < 0.05$) , except for /y/, /uw/, /d/, /v/, /D/, /J/, /tS/ and /dZ/. In addition we observe a larger difference between the vowels of BN and BC.

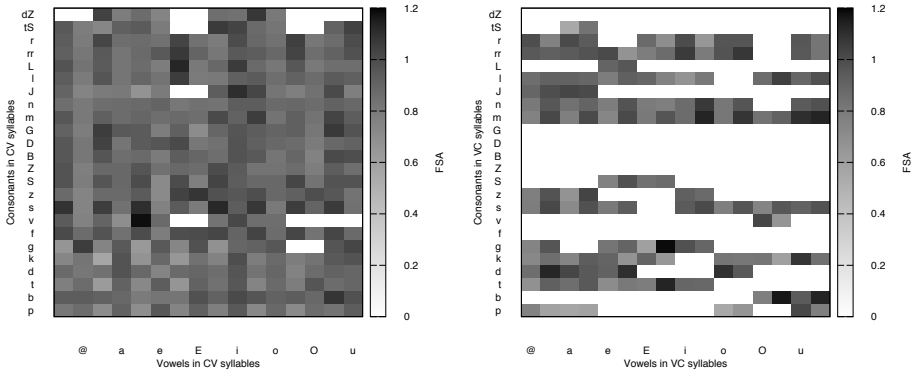
Reconsidering the assumption that duration reduction may not necessarily trigger phonetic acoustic reduction, we measured the Spearman correlation ρ (monotone functional relationship) between corresponding samples of duration and FSA of a phonemic segment. Apart from the phonemes /uw/, /w/, /Z/, /m/, /rr/ and /dZ/ that did not show a significant correlation ρ, all other phonemes indicated small but statistically significant Spearman correlation ($p < 0.05$). As a conclusion, reduction in duration may not necessarily involve a reduction in feature space.

Because vowel reduction was also considered a result of contextual coarticulation and assimilation processes, we conditioned the FSA evaluation of a particular phonemic segment to its context, and constrained the evaluation to syllables as the phonological units that particularly tie a sequence of phonemes.

Although this contextual FSA has been carried out on BC and BN speech data, and specific to CV, VC, CVC and VCV syllables, we confine subsequent description to the findings in CV and VC syllables of BC speech data. We noticed similar FSA characteristics of CV and VC syllables in CV and VC segments of CVC and VCV syllables.

The CVC and VCV syllables were evaluated solely on their CV and VC segments.

Figure 3 depicts the values of the FSA of phonemes (a subset of all vowels and consonants) in CV and VC syllables. White tiles indicate a missing occurrence of the corresponding syllable. The vowels at the column bottom address the FSA of the vowels given the consonants that are addressed by the row.



(a) Feature Space Analysis in CV Syllables

(b) Feature Space Analysis in VC Syllables

Fig. 3: Feature Space Analysis in CV and VC syllables.

Evaluating the FSA of CV syllables in Figure 3a allows for several observations: In summary, the FSA values of neighbouring phonemes in almost all syllables are different, most of them significantly (ANOVA $p < 0.05$). Plosives in CV syllables exhibit a significantly lower FSA value indicating a stronger acoustic reduction than their succeeding vowels (ANOVA $p < 0.05$), except for 4 syllables (3 of them having a significantly higher FSA value). Fricatives generally show a strong resistance against reduction compared to other consonants. Their FSA value is significantly higher than of most of their succeeding vowels except for the allophonic fricatives /B,D,G/ with succeeding vowel /i/ where they tend to the contrary. Nasals show an ambivalent context-conditioned FSA. They exhibit less reduction than the succeeding central vowel /@/ respectively a significantly higher FSA value but more significant reduction than their succeeding front vowels. Contrary to these findings, the close back vowels that show more reduction if the difference is significant (ANOVA $p < 0.05$). The liquids expose identical tendency of the difference in reduction for the same succeeding vowel (when difference is significant, $p < 0.05$), but ambivalent relationship for different succeeding vowels. The FSA of affricates /tS/ and /dZ/ in context often suffers from an insufficient number of samples, and since the joined phenomena of plosive onset and a fricative offset would require a rather distinct segmentation, we avoid a particular conclusion.

The FSA of phonemes in VC syllables in Figure 3b exposes differences between vowels and their succeeding consonants, many of them statistically significant. Most plosives and fricatives exhibiting higher FSA values than their preceding vowels, almost all of those significant (ANOVA $p < 0.05$). All nasals show a significantly higher FSA value than their preceding vowels (ANOVA $p < 0.05$). Lateral liquids suggest a tendency towards less reduction than their preceding vowels (though only in a few of them significantly), while both liquid flap and trill expose the opposite. Affricates in VC syllables are not representa-

tive. As an essential observation we note that plosives have a larger FSA than their preceding vowels in VC syllables in contrast to CV syllables where plosives have a smaller FSA.

4 Discussion

The FSA provides a common means to analyse and compare the degree of feature space reduction not only on average between corpora of different speaking styles, but also and in particular to assess the degree of reduction between phonemes and phonemes given a particular context.

Studies on coarticulation and assimilation in Catalan suggested the degree of articulatory constraint (DAC) model [14][15]. It deals with the direction of coarticulatory effects in CV and VC sequences. More specifically, it asserts anticipatory and carryover effects. The model assigns a DAC to every phoneme depending on its tongue dorsum constraint during the sound production, effectively qualifying to what degree a particular phoneme is constrained and resists coarticulation. For instance, fricatives are highly resistant to coarticulation in terms of the DAC model. Vowels receive different DAC. Supportive to these assumptions are the observations, that the FSA values of fricatives in CV and VC are consistently higher than their neighbouring vowels.

The assertions of the DAC model also eliminate substantial effects of stress and a role of syllable affiliation on coarticulation. The latter would support the observation that the FSA characteristics in CV or VC syllables are similar to those of corresponding CV or VC segments in other syllables. But it would question the finding that plosives have a larger FSA value respectively are less reduced than their preceding vowels in VC syllables in contrast to CV syllables where plosives have a smaller FSA value.

5 Conclusions

We rephrased and applied a phonemic segment specific FSA, and evaluated two different Catalan spoken language databases of distinct speaking styles. The evaluation of BC and BN speech data show a significant reduction in duration and in terms of the feature space spanned between phonemes exhibited. We observed minimal but significant correlation between the size of the feature space and duration of a phoneme. The context dependent analysis of the feature space reduction showed significant relations between neighbouring vowels and consonants in CV, VC and CVC syllables. The comparison between CV and VC syllables indicates that the reduction of plosives depends on the syllable position - initial versus final. Furthermore context specific analysis also shows that a more reduced phoneme may be observed next to a less reduced phoneme. The observations have been discussed in conjunction with the DAC model. The FSA is expected to facilitate a refined modelling in speech synthesis and recognition. In current ongoing work we explore the induction of the FSA as acoustic reduction measurement into AM for ASR.

Acknowledgement: Above described research was supported by the Centro para el Desarrollo Tecnolgico Industrial (CDTI) - project Cenit 2009 BUSCAMEDIA CEN-20091026, by the Spanish government - project BUCEADOR TEC2009-14094-C00, and by the FI-DGR-2009 grant of the Catalan government. We would like to thank Televisió de Catalunya (TVC), the public Catalan broadcaster, for their collaboration.

References

1. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis. IEICE - Trans. Inf. Syst. **E88-D**(3) (March 2005) 502–509
2. Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A.: Effect of Speaking Style on LVCSR Performance. In: Proc. ICSLP. (1996) Addendum, 16–19
3. Furui, S., Nakamura, M., Ichiba, T., Iwano, K.: Why Is the Recognition of Spontaneous Speech so Hard? In: Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence, Karlovy Vary, Czech Republic, Springer (2005) 9–22
4. Lindblom, B.: Spectrographic study of vowel reduction. Acoustical Society of America Journal **35**(11) (nov 1963) 1773–1781
5. Pols, L.C.W., van Son, R.J.J.H.: Acoustics and perception of dynamic vowel segments. Speech Communication **13**(1) (1993) 135–147
6. Moon, S.J., Lindblom, B.: Interaction between duration, context, and speaking style in english stressed vowels. Acoustical Society of America Journal **96**(1) (1994) 40–55
7. van Son, R.J.J.H., Pols, L.C.W.: An acoustic profile of consonant reduction. Proc. of ICSLP'96 **3** (1996) 1529–1532
8. van Bergem, D.R.: Acoustic vowel reduction as a function of sentence accent. Speech Communication **12** (1993) 1–23
9. Herrick, D.: An Acoustic Analysis of Phonological Vowel Reduction In Six Varieties of Catalan. PhD thesis, University of California, Santa Cruz (September 2003)
10. Wrede, B.: Modelling the Effects of Speech Rate Variation for Automatic Speech Recognition. PhD thesis, University of Bielefeld, Bielefeld, Germany (June 2002)
11. Schulz, H., Fonollosa, J.A.R.: A Catalan Broadcast Conversational Speech Database. In: Speech and Language Technologies for Iberian Languages. (September 2009) 27–30
12. Pachès, P., de la Mota, C., Riera, M., Perea, M.P., Febrer, A., Estruch, M., Garrido, J.M., Machuca, M.J., Ríos, A., Llisterri, J., Esquerra, I., Hernando, J., Padrell, J., Nadeu, C.: Segre: An automatic tool for grapheme-to-allophone transcription in Catalan. In: Proc. of Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC, Atenas (2000) 52–61
13. Carbonell, J., Llisterri, J.: Illustrations of the IPA: Catalan. Journal of the International Phonetic Association **22** (1992) 53–56
14. Recasens, D.: Coarticulatory patterns and degree of coarticulatory resistance in Catalan CV sequences. Language and Speech **28** (1985) 97–114
15. Recasens, D., Pallares, M.D.: Coarticulation, assimilation and blending in Catalan consonant clusters. Journal of Phonetics **29** (2001) 273–301

A preliminary approach to forensic speaker recognition using phonemes

Pedro Univaso[1], Miguel Martínez Soler[1,2], Diego Evin[1], and Jorge Gurlekian[1]

[1] Laboratorio de Investigaciones Sensoriales, INIGEM, UBA-CONICET
Córdoba 2351 Piso 9 Sala 2, Buenos Aires, Argentina
punivaso@yahoo.com.ar,{miguelmsoler,diegoevin}@gmail.com,jag@fmed.uba.ar
http://www.lis.secyt.gov.ar/
[2] Facultad de Ingeniería, Universidad Austral,
Av. Garay 125, C1063ABB Buenos Aires, Argentina

Abstract. Present work focus on speaker identification using phonemic information. An ASR system -based on HMMs- is employed to extract acoustic information from phonemes as speaker identity features. Main contribution resides on proposing phoneme forced alignment which could be performed at the forensic environment. This approach with manual intervention, is further combined with a GMM approach, resulting in an expected improvement on performance. Tests on a fixed-phone database of 136 Argentine-Spanish speakers were performed to evaluate the proposed approach in two conditions: first, using the same recording session and channel for both the suspect and evidence and second adding babble noise to the evidence. Results for the ideal condition show an error rate reduction of 68% relative to the GMM baseline system. The difference in favor of HMM with forced alignment against GMM holds for different SNR conditions.

Keywords: Automatic speaker recognition, Hidden Markov Models (HMM), Gaussian Mixtures Models (GMM), Logistic Regression, Support Vector Machines (SVM).

1 Introduction

Speaker recognition main application fields are found in forensics and in security systems. Speaker identification as seen in forensics is initiated from voice recordings produced at a criminal situation. These recordings are named dubitable or evidence, which are later matched with recordings called indubitable or suspicious that belong to a known person. Standard techniques employed by forensic recognition experts are based on those known as speaker verification by speech scientists. These techniques propose a Bayesian approach to solve a likelihood proportion between matching hypothesis. Some authors [1] claim that forensic situations must be solved using a robust method that extends this approach by holding the principle of innocence of the suspect candidate. Others [2] indicate to change the standard verification binary decision of acceptance or refusal for the a

posteriori probability of the hypotheses, due to the high environment variability and to several factors that affect forensic recognition [3].

In order to represent each speaker voice, generic acoustic probabilities based on GMMs [4] are employed in actual state of art speaker recognition systems. During their evolution, the use of distinctive phonetic features in characteristic vectors can be found in some previous work. Some of them started by using phoneme based models represented by HMMs [5]. Experiments presented by Kajarekar et al. [6] arranged clusters of wide phonetic categories to produce each speaker models.

Nevertheless, this approach was discarded both in the academic community and the application developments when contrasted with the performance obtained with GMM based systems, because of practical advantages of GMM solutions, as their computational efficiency and text independence of samples.

A drawback of GMMs system implementations is that they usually do not take advantage of linguistic information both phonological and phonetic present in running speech. Phoneme discriminative power has been studied by several authors [7, 8] but the way for incorporating this knowledge in speaker recognition systems still is an open issue. High level information of prosodic features as intonation, rhythm and accent are not adequately exploited. Hansen et al. [9] provided a useful background to the present work as they showed improvements relative to standard GMM approach by using a GMM for each phoneme. Some investigations from Stanford Research Institute evaluated the introduction of high level features in GMM based systems. It is worth to mention two works recently presented by Stolcke et al. [10] and Kajarekar [11] who added words, syllables and phonemes as parameters for the speaker recognition task. SRI systems showed the best performance during the NIST evaluation contest. Systems developed at SRI as well as other systems which used models based on high level features [12] produce a final model as a combination between both high and low level models. The result is always an improvement relative to the isolated model. Model combination is achieved in different ways: by neural networks, support vector machines, logistic regression or some other particular classifiers [13].

Inspired by this successful approach, we present a new contribution on speaker recognition at the forensic domain that evaluates the use of acoustic information present at phoneme level. Our proposal is to produce a more precise phoneme model using forced alignments in a HMM. This can be done because at the forensic domain we can produce almost perfect segmentation and labelling to catch later on phoneme information in a stochastic way by mean of HMMs. Our results are also compared with a GMM based system doing the same job. Finally a combination of both systems is obtained by a logistic regression classifier. A linear combination classifier is also employed for this purpose.

The rest of this paper is as follows: Section 2 presents the proposed method based on phonological information and its relation with a standard automatic method. Section 3 shows the strategy of evaluation. Section 4 presents results obtained, which are discussed in section 5, to finally present our conclusions and future work in sections 6 and 7, respectively.

2 Methodology

The proposed methodology consists in the extraction of acoustic information from phonemes by forced alignment performed with an ASR system. The motivation for doing this is that a human can make word and phonetic transcripts in a forensic environment, both for the known and unknown speaker samples that are going to be tested. This approach with manual intervention, is further combined with a Gaussian mixture automatic recognition method, resulting in an additional improvement on performance.

GMM Model. Our GMM model is based on the Universal Background Model (UBM) presented in [14]. Bayesian adaptation via maximum a posteriori estimation (MAP) is used to adapt speaker models from the UBM. A simplified schematic of the speaker recognition system based on the methodology GMM-UBM can be seen in Fig. 1.

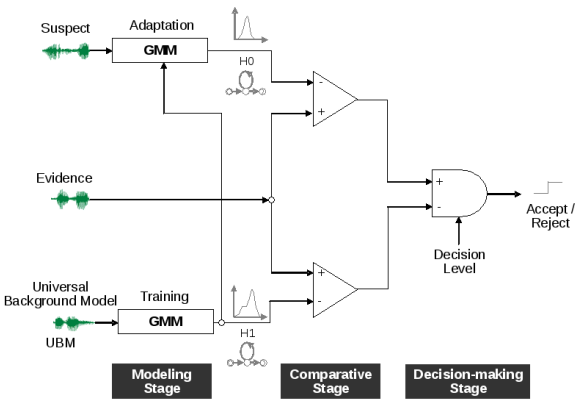


Fig. 1. GMM-UBM methodology.

Hidden Markov Models Fig. 2 shows the procedure used in this study to recognize speakers using an HMM-based ASR system. The generation of the universal reference (UBM) speakers acoustic models was performed through a training process in stages according to the methodology proposed by Young [15] for ASR systems. HMMs are usually built to model either single isolated phones or context-dependent triphones. The latter is the approach used in this work, so from now on when we refer to a phone in an HMM model we are talking about one of the three states of the HMM triphone model.

Adapting the model of the suspect was conducted similarly to the case of GMM, re-estimating each HMM corresponding to UBM with information of the suspect. The step of automatic speech recognition is performed using a general

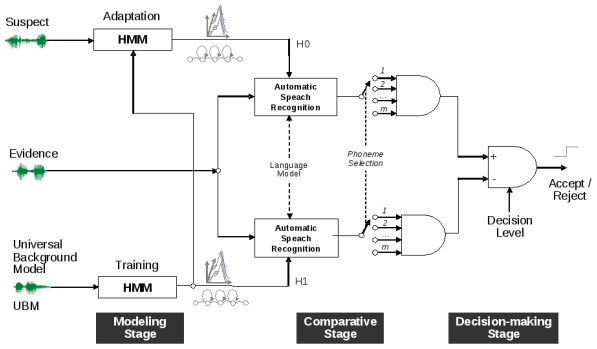


Fig. 2. HMM-UBM methodology.

purpose word recognizer based on the Viterbi algorithm. An alternative to the above method is to employ a forced alignment, instead of the word recognition, for which the Viterbi algorithm uses a word network to recognize the corresponding transcriptions. This alternative is more accurate and fast and does not require the generation of a language model, but requires a human-made transcript.

In the case of forensic applications generally we have these transcripts of the evidence and this approach is to be used extensively in this work. In the final stage of comparative analysis, we calculate the average of the likelihoods of each of the phonemes that were recognized to finally obtain an overall average value for each of the comparisons. These two values are representative of the similarity between the evidence and the UBM and between evidence and the suspect, which are used in the final decision analysis using a likelihood ratio test.

Combined methods. For the first combined methodology proposed here an SVM logistic regression model were trained by fitting logistic regression models to the outputs of the support vector machine. The input data for the SVM were obtained at the output stage of GMM and HMM methods, expressed in the formulas (3) and (8) respectively. For SVM training, the Weka implementation of the SMO algorithm [18] were used.

The second approach is a linear combination of the outputs of both systemas, weighted by an α factor in the interval $[0; 1]$ for the HMM system output score, while its complement si applied to the GMM output, according to:

3 Evaluation

3.1 Database

The database used, is part of the SALA I Project (SpeechDat Across Latin America) [16]. The style of speaking corresponds to read paragraphs taken from newspapers and books of Argentina or developed by linguists. Recordings were

made through the fixed telephone network through a computer equipped with an AVM-ISDN-A1 board and a basic access interface ISDN (BRI).

For this work we selected sentences of continuous speech from the SALA database Argentina, South region, which will be called SALA I-Continuous. This region includes the provinces of Buenos Aires, Santa Fe, Entre Ríos, La Pampa, Neuquén, Rio Negro, Chubut, Santa Cruz and Tierra del Fuego. The corpus was delimited to 1,301 words, with a total of 9,948 words, corresponding to a vocabulary of 2,722 different words, issued by 136 speakers (47 males and 89 females) for 99 minutes of recording. For final comparisons between the three speaker recognition systems, a 10-fold cross-validation method was used. For each partition we used the data from 124 speakers for the formation of UBM and 12 speakers for testing identities, testing 120 speaker in total due to the cross-validation schema.

The average duration of the recordings for each speaker was 44 seconds. We used 2 segments of the testing speakers (5 seconds on average) as unknown speaker samples while the remaining utterances were used to adapt speaker models. The cartesian product between test segments and speaker models yields 24 target and 264 non-target tests per fold, totalizing 288 cases per fold and 2880 for the whole experiment. While the paucity of evidence can lead to low recognition rates, it also allows to represent one of the real drawbacks of current forensic field. Using data from women only, improvements were of 54%, but due to the small amount of data, final experiments were performed using the complete database.

3.2 Reference system

HMM models were built using a tool developed by the University of Cambridge: HTK Toolkit ver. 3.4 [15], free for academic use. Digitizing the audio signal is performed at a sampling frequency of 8 kHz and 16 bits, signal mean substraction to eliminate any offset from the analog recording stage. We employed a 25 ms Hamming window at a 10 ms rate, using a pre-emphasis filter with coefficient 0.97, having normalized energy of the sentence and using a logarithmic scale for energy.

The parameters used for creating the models were 13 MFCC coefficients: Mel-Frequency Cepstral Coefficients, including zero coefficient whose value is approximately proportional to the fundamental frequency F0, to which were added delta and acceleration transformations, forming a total of 39 parameters. The selection of the phoneme as the unit of speech to be used in this work was based on the phonetic alphabet SAMPA Speech Assessment Methods: Phonetic Alphabet, adapted for Spanish in Argentina [17]. In addition to the 30 units of SAMPA phonetic alphabet formed with the 22 phonemes of Spanish and 8 allophones often used in Argentina, a model of silence generally present between phrases was included, which is associated with a model of short pause between words, completing a total of 31 monophones that represent the speech of Spanish in Argentina. HMM was created as a 3-state left to right for each of the 31 monophones, except the short pause that is associated with the central

state silence model. The construction of context-dependent models (triphones) was performed by grouping monophones by a clustering based on phonetic decision trees, from which we obtained a total of 1,010 clusters. Eventually older models were transformed in associated Gaussian mixture models, designated in the literature as semi-continuous models (SC-HMM) using a splitting method to achieve a number of 256 mixtures by model. The best fit is achieved with only one re-estimation, increasing the speaker recognition error by 41% when using 2 re-estimates.

4 Results

Table 1 shows comparative results of equal error rate (EER) and minimum detection cost ($C_{Det-min}$), produced by the baseline method and the proposed method with different combination procedures. False acceptances and false rejections are often called false alarms and misses, respectively, and this is the convention that we will use in the rest of this paper. C_{Det} metric is widely used by the speaker recognition community and it is a cost function defined as a weighted sum of miss and false alarm error probabilities. Lastly, $C_{Det-min}$ is the minimum detection cost that can be achieved by using the detection thresholds that minimize the detection cost.

Table 1. Average equal rate error and minimum cost rate in 10-fold cross-validation for each tested method. Additional experiments under several SNR conditions were performed using GMM and HMM forced alignment.

SNR		GMM	HMM word recognizer	HMM forced alignment	SVM Log. Regression	Linear Comb. EER opt.	Linear Comb. $C_{Det-min}$ opt.
EER	27dB (Original)	13.4	15.5	4.5	5.9	4.2	4.2
	20dB	37.3	-	17.8	-	-	-
	15dB	42.9	-	20.6	-	-	-
	10dB	46.3	-	28.4	-	-	-
	5dB	47.2	-	34.1	-	-	-
	0dB	48.9	-	41.0	-	-	-
$C_{Det-min}$	27dB (Original)	8.7	6.7	2.0	1.9	2.0	1.8
	20dB	10.0	-	6.0	-	-	-
	15dB	9.9	-	7.6	-	-	-
	10dB	10.1	-	8.6	-	-	-
	5dB	10.0	-	9.4	-	-	-
	0dB	12.2	-	10.0	-	-	-

As using the same recording session and channel both for known and unknown identity speech segments is far from real forensic conditions, we added babble noise[3] to the unknown speaker samples at several different SNR levels (20dB, 15dB, 10dB, 5dB, and 0dB). This forced a mismatch between the training and testing conditions, making the identification task more challenging. The use of

[3] http://spib.rice.edu/spib/data/signals/noise/babble.html

babble noise suppose a great challenge, owing to the fact that the speech signal and this type of noise are very similar.

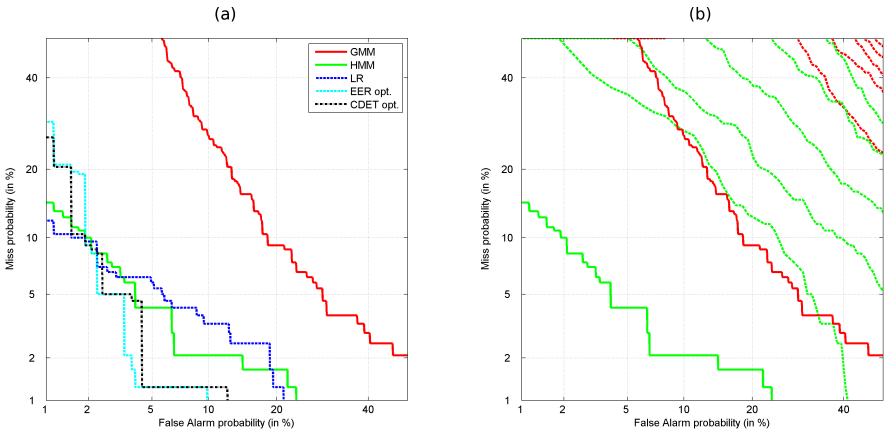


Fig. 3. *(a)* Average DET curves over the 10 folds for GMM, HMM with forced alignment, logistic regression (LR), and both linear combinations (optimizing EER and $C_{Det-min}$, respectively). *(b)* Average DET curves for GMM and HMM with forced alignment. The dashed lines correspond to several SNR, after babble noise is added. For each method 0dB, 5dB, 10dB, 15dB and 20dB curves appear in sequence, where greater SNR values correspond to the curves nearer to the origin.

In order to get the SNR value for the original condition (before adding noise), a voice activity detection algorithm was used to calculate rms powers of the speech and non-speech part of each utterance. SNR values were computed then as the log ratio between the rms power of the speech part (minus the rms power of the noise part, since the noise is also present in speech segments) and the rms power of the non-speech part. The average SNR on the corpus before noise addition was 27dB.

The HMM recognition methodology produced more errors than the GMM baseline, as expected. The forced alignments in HMMs introduced the most significant change with an improvement of 66% in EER, and 77% in minimum-cost rate related to GMM. Then, various combination strategies were built and tested based on HMM-forced alignment and GMM. Parameters required for the logistic regression on SVM were determined by a previous run on the training data. The linear combination optimizing EER was achieved with a factor α=0.85, and the optimization of $C_{Det-min}$ was obtained with α=0.9. GMM-HMM combinations, using both logistic regression and a linear combination algorithm, achieve further improvements of 1% and 2% respectively. Linear combination α parameter and logistic regression (with SVM) were estimated by training on the same 10-fold

cross-validation scheme that was used for GMM and HMM, that is, using for training the folds that were not being used for testing in each case.

In the mismatched condition, after noise addition, the differences between HMM forced alignment and GMM persisted. In both 20dB and 15dB SNR condition, particulary, it seems that the difference in favor of HMM forced alignment gets wider, al least in terms of EER. The DET plots in Fig. 3 (b) support these results and show the performance of both methods at different operating points.

DET (detection error trade-off) curves in Fig. 3 (a) show overall results from the experiments for GMM, HMM forced alignmen, and combinations. The overall differences of the tested methods can be seen in the displacements of the curves. However, average EER and $C_{Det-min}$ can not be plotted here, since their averages over 10 folds are not necessarily equal to the same metrics calculated on the average curves.

5 Discussion

Results of HMM forced alignment and its combinations show the best performance between all experiments we made. The use of forced alignments has proved in this work, as it could be supposed in advance, a clear improvement over both GMM and HMM recognition. Considering that the scientific community has been working mainly on automatic speaker recognition tasks, the forensic field could benefit from this basic idea of using the known text to produce forced alignments and then to continue with the automatic recognition paradigm using a universal background model of a language region.

In the ideal condition of using the same recording session and channel for both the suspect and evidence voices, results show an error rate reduction of 68% relative to the GMM baseline system. The next step would be to test whether this improvement remains after using speech from different sessions and channels. In this respect, our results on SNR mismatch conditions suggest that force alignment could give some robustness when speech samples to be tested present these kind of problems. We must notice, however, that we have evaluated the methods at fixed operating points (corresponding to EER and $C_{Det-min}$). In order to get a good assessment for a forensic setup we should consider a likelihood ratio output in the future.

In summary, the semiautomatic approach with forced alignment presents results that are highly promising and should be tested on bigger corpora having different sessions and channels for each speaker.

6 Conclusions

Determining the most efficient way to use speaker intrinsic information to improve speaker identification is still a debt we have today. The methodology presented here is an alternative approach that attempts to contribute to solve this issue. The way phonemes are produced by the speaker and their variations are highly speaker dependent.

Modeling speaker voice using HMM contributes to enhance phonological information present at phonemic level. The use of forced alignment in HMMs is possible for forensic purposes, where the court requests the use of speaker identification techniques using transcriptions made in advance. System combinations produce a further improvement, confirming what was stated in other studies [10], and where the linear combination methodology is simpler and more effective than logistic regression considering EER metric.

7 Future Work

First, increasing the amount of training data by using the full SALA database will permit to adequately assess performance with triphones, given that there is a compromise between the quality of representation of the context and number of samples available to adequately estimate models for each of these contexts. We are also building up a new multi-channel and multi-session database of Argentine spanish that will be used to test the method in channel and session mismatch conditions. Also we expect to correlate speaker profiles with gender, age, dialect, type of telephone channel and environment situation. Other methodologies will be analyzed for adaptation (MLLR, CMLLR) instead of the re-estimation used, based on Baum-Welch algorithm.

We expect to evaluate the role of individual phonemes. Giving them differential weights inspired in perceptual judgments and in previous work [8] where phonetic groups as nasals, vowels and fricatives have already demonstrated to produce better results in speaker discrimination. Also, the proposed method will permit research in a flexible manner on the influence of other parameters that contribute to speaker recognition, such as: phoneme durations and silences, distribution of phonemes and words used in spontaneous conversations.

Forensic approach should also consider other metrics like C_{llr} related to the adjustment of a reported conditional probability instead of a hard decision response. C_{llr} is a application-independant information theoretic measure that considers how well all scores represent the likelihood ratio and that penalizes for errors in score calibration. We plan to use this metric in the future, as well as Tippett and APE plots [19].

Acknowledgments. To project PID 35891 "Development of techniques for speaker recognition", awarded by the Ministry of Science, Technology and Productive Innovation of Argentina.

References

1. Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., Ortega-Garcia, J.: Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. IEEE Transactions on Audio, Speech and Language Processing, 15(7), 2104-2115 (2007)

2. Campbell, W., Reynolds, D., Campbell, J., Brady, K.: Estimating and evaluating confidence for forensic speaker recognition. In: Proc. ICASSP, pp. 717-720, Philadelphia (2005)
3. Campbell, J., Shen, W., Campbell, W., Schwartz, R., Bonastre, J., Matrouf, D.: Forensic speaker recognition. IEEE Signal Processing Magazine, 26(2), 95-103 (2009)
4. Reynolds, D. A.: Speaker Identification and Verification Using Gaussian Mixture Models. Speech Commun., 17, 91-108 (1995)
5. Matsui, T., Furui, S.: Concatenated Phoneme Models for Text-Variable Speaker Recognition. In: Proc. of ICASSP-93, Vol. 2, pp. 391-394, Minneapolis (1993)
6. Kajarekar, S. S., Hermansky, H.: Speaker Verification Base on Broad Phonetic Categories. In: Proceedings of Speaker Odyssey 2001, pp. 201-206, Crete, Greece (2001)
7. Sambur, M. R.: Selection of Acoustic Features for Speaker Identification. Acoustics, Speech and Signal Processing, IEEE Transactions on, 23(2), 176-182 (1975)
8. Eatock, J. P., Mason, J. S.: A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes. In: Proceedings of ICASSP 1994, Vol. 1, pp. 133-136, Adelaide, Australia (1994)
9. Hansen, E., Slyh, R., Anderson, T.: Speaker Recognition using Phoneme-Specific GMMs. In: Proceedings of Speaker Odyssey 2004, pp. 179-184, Toledo, Spain (2004)
10. Stolcke A., Shriberg E., Ferrer L., Kajarekar S., Sonmez K., Tur G.: Speech Recognition as Feature Extraction for Speaker Recognition. In: Proc. SAFE 2007: Workshop on SP Applications for Public Security and Forensics, pp. 39-43, Washington, D.C. (2007)
11. Kajarekar S., Scheffer N., Graciarena M., Shriberg E., Stolcke A., Ferrer L., Bocklet T.: The SRI NIST 2008 Speaker Recognition Evaluation System. In: Proc. ICASSP, pp. 4205-4208, Taipei (2010)
12. Campbell, J. P., Reynolds D. A., Dunn, R. B.: Fusing High- and Low-Level Features for Speaker Recognition. In: Proceedings of Eurospeech 2003, pp. 2665-2668 Geneva, Switzerland (2003)
13. Ferrer, L., Graciarena M., Zymnis, A., Shriberg, E.: System Combination using Auxiliary Information for Speaker Verification. In: Proc. ICASSP, pp. 4853-4856, Las Vegas (2008)
14. Reynolds D., Quatieri T., Dunn R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41 (2000)
15. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtech, V. Wooland, P.: The HTK Book. Cambridge University Press (2006)
16. Moreno A.: SALA: SpeechDat Across Latin America. In: Proceedings of the I Workshop on Very Large Databases, Athens, Greece (2000)
17. Gurlekian, J., Colantoni, L., Torres, H., Rincn, A., Moreno A., Mario J.: Database for an automatic Speech Recognition System for Argentine Spanish. In: Proc. Of the IRCS Workshop on Linguistic databases, pp. 92-98, Pennsylvania (2001)
18. Platt, J.: Machines using Sequential Minimal Optimization. In: B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, (1998)
19. Brummer, N., du Preez, J.: Application-Independent Evaluation of Speaker Detection Comput. Speech Lang., 20(2-3), 230-275 (2006)

Spectral Envelope Representation using Sums of Gaussians

Anderson Fraiha Machado[1], Antonio Bonafonte[2], and Marcelo Queiroz[1]

[1] Institute of Mathematics and Statistics, University of São Paulo, São Paulo – Brazil,
dandy@ime.usp.br, mqz@ime.usp.br,
[2] Universitat Politècnica de Catalunya, Barcelona – Spain,
antonio.bonafonte@upc.edu

Abstract. The aim of this paper is to present a new approach to spectral representation using sums of Gaussian distributions. Sums of Gaussians provide an intuitive representation for frequency bands of a signal spectrum as well as formant regions. The representation of spectral envelopes using Gaussian parameters $\{a, \mu, \sigma\}$ simplifies the expression of important tasks such as frequency warping and formant manipulation. Marquardt's algorithm has been extended to estimate parameters of the Gaussian models for each frequency band, allowing each Gaussian parameter to be either optimized for fitting a given spectral sub-band, or else have a fixed value for reducing the number of model parameters. This allows for several choices on the sets of free/fixed parameters and the sizes of models. Experimental results show that the models proposed offer an accurate approximation of spectral envelope, and provide good perceptual results when applied to pitch shifting.

Keywords: spectral envelope representation, sum of Gaussians

1 Introduction

The estimation of spectral envelopes is one of the most important and well-established topics in Signal Processing in general, and in Speech Technology in particular. Several well-known applications, such as speech recognition, speaker verification, statistical speech synthesis, diarization and voice conversion, require obtaining a faithful representation of the spectrum envelope using few parameters. The Mel-Frequency Cepstrum Coefficients (MFCC) are the most widely-used spectral envelope representation. In statistical speech synthesis LPC-based cepstrum coefficients are also frequently used. In speech recognition, we are interested in obtaining a representation which is speaker-independent, or at least one that can be easily transformed into a speaker-independent representation. For instance, vocal-tract length normalization (VTLN) has been successfully included in the adaptation of MFCC to reduce speaker variability and simplify speech recognition tasks. VTLN has also be used in voice conversion to transform the spectrum of a given acoustic signal from one speaker to another. Formant-based

representations are an alternative which allows one to characterize a spectrum envelope using few parameters. However, this representation is not accurate enough and is not easily estimated under all conditions.

Gaussian-Mixture Models (GMM) are frequently used in speech technology to represent the probability density function (pdf) of speech parameters (e.g. MFCC of one particular speaker) [1]. Even if the pdf does not strictly fit a given model, it provides accurate representation for most practical situations, given that enough Gaussian components are included in the mixture. Well-known algorithms are widely-used to estimate the parameters of the pdf from training data (e.g: EM for maximum likelihood estimation). Many techniques have been proposed to adapt or transform the sum of Gaussians representing one acoustic signal class to another, (e.g: voice gender conversion [2] and voice conversion [3]).

In this paper we introduce a parametric spectral estimation method that assumes that the spectral envelope can be represented as a sum of Gaussians. Notice that the function to be represented – the spectral envelope – depends on only one dimension (the frequency) and is strictly positive. Moreover, it is expected that parameters of each Gaussian component (center frequency, amplitude and bandwidth) reflect temporal continuity between adjacent frames.

Several *models* are proposed in Section 2 with varying constraints, and estimation methods are derived. Section 3 presents an application framework for pitch shifting, which is a required step in statistical speech synthesis (and in some unit-selection speech-synthesis systems). Finally, Section 4 presents both objective and subjective evaluations of the quality of the transformed signal.

2 Gaussian Spectral Fitting

The problem of fitting a sum of Gaussians to a spectral representation of a signal has been addressed in previous work. In 1996, Zolfaghari [1] presented a method for formant estimation using mixtures of Gaussians based on the well-known EM Algorithm.Also, the same author has proposed a Bayesian modelling of the STRAIGHT spectral envelope [4] using mixtures of Gaussians [5]. Although these methods provide good results in representing formant regions, they are not able to model spectral nuances that are required in speech reconstruction.

A recent work [6] has also proposed a spectral envelope model using sum of Gaussians in which the amplitude and central frequency of each component is defined based on each local peak of the spectrum. The variance estimation considers the distance between neighboring peaks, and does not consider the overlaps between Gaussians in the mixture.

Unlike these methods, in what follows uses pre-defined frequency bands to get a optimal Gaussian parameters using a adaptation of Marquardt's algorithm.

The sum of L Gaussians is defined as $G(a, \mu, \sigma) = \sum_{k=1}^{L} G(a_k, \mu_k, \sigma_k)$ with each Gaussian $G(a_k, \mu_k, \sigma_k) : \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$[G(a_k, \mu_k, \sigma_k)](f) = a_k \, e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}}, \tag{1}$$

where a_k, μ_k and σ_k are, respectively, the amplitude, central frequency and standard deviation of the k-th Gaussian component. The modelling problem is to find an approximation of a function $S(f)$ by the above sum of Gaussians in such a way that the estimation error is minimal.

In this work, $S(f)$ represents the envelope of the DFT of a discrete signal; for instance, $S(f)$ can be derived using a non-parametric representation of the spectral envelope. This fitting problem is equivalent to solving a system of nonlinear equations with $3L$ coefficients using a nonlinear least squares estimator. Furthermore, since the spectral envelope is assumed to have positive values, it becomes necessary to use an estimation method that constrains $a_k > 0$ for all k, which is addressed below.

2.1 Sequential Gaussian Fitting Algorithms

In 1994, Goshtasby and O'Neill [7] proposed a method to fit a curve with a sum of Gaussians based on the algorithm of Marquardt, whose initialization is determined from derivative values and zero-crossing rates. Although this proposed system obtains an accurate approximation,it is not applicable to the problem of spectral representation, since it allows the use of negative amplitudes.

This paper presents a variant of Marquardt's algorithm [8] constraining $a_k > 0$ for all k. The original proposal [7] estimates all Gaussian parameters (a_k, μ_k, σ_k) $\forall k$ in parallel, making it inappropriate for controlling specific parameters of each Gaussian, such as the value of a_k. Therefore we propose a modification where we treat separately each Gaussian component $G(a_k, \mu_k, \sigma_k)$, which fits a pre-specified section of the spectrum defined by a spectral sub-band S_k. There are two important steps of the proposed algorithm that are intertwined. The first one is the definition of the spectral sub-bands S_k, which is addressed below, and the second is the parameter estimation for each Gaussian $G(a_k, \mu_k, \sigma_k)$, so that it fits the respective spectral sub-band S_k, which is addressed afterwards.

Defining spectral sub-bands In some applications, particularly in speech processing (e.g. such as statistic synthesis [9] and voice conversion [3]), it is often desired that the Gaussian coefficients are estimated in fixed frequency bands, specially the means of each Gaussian component.

Given a set of L frequency bands centered on c_k, we first split the spectrum in L band-limited portions S_k, that will be fitted to a Gaussian component in the sum of Gaussians. Each S_k is defined by windowing the spectral signal with a window W_k, usually a triangular window, so that $\sum_k S_k = S$. The set of sub-bands used in this work are the critical bands of the Bark scale, represented as a set of windows B_k.

Algorithm 1.1 describes how each sub-band S_k is obtained and initially modelled by a static Gaussian, with its parameters obtained directly from S_k. The choice of initial Gaussian parameters uses the global peak $p = (x, y)$ of each spectral sub-band S_k, reestimated with parabolic interpolation; in this work, a_k^0, μ_k^0 and σ_k^0 are initialized as y, x and width of S_k, respectively.

Algorithm 1.1 (a^*, μ^*, σ^*) = `Band_Gauss_Estimation(`S`,`L`,`B`,`ϵ`)`
1 **for** k = 1 to L
2 $S_k \leftarrow S \cdot B_k$
3 $a_k^0 \leftarrow \max(S_k)$ /* using quadratic interpolation of amplitude */
4 $\mu_k^0 \leftarrow \arg\max(S_k)$ /* using quadratic interpolation of frequency */
5 $\sigma_k^0 \leftarrow$ bandwidth(S_k)
6 $(a_k^*, \mu_k^*, \sigma_k^*) \leftarrow$`Gaussian_Fitting(`$S_k$`,`$(a_k^0, \mu_k^0, \sigma_k^0)$`,`$\epsilon$`)`
7 **end for**

Each Gaussian component $G(a_k, \mu_k, \sigma_k)$ is then optimally fitted to the spectral sub-band S_k using Algorithm 1.2 `Gaussian_Fitting`, which will be discussed next.

Optimally fitting each Gaussian component Given the k-th Gaussian, the goal of Algorithm 1.2 `Gaussian_Fitting` is at each iteration to update the parameters (a_k, μ_k, σ_k) by making $(a'_k, \mu'_k, \sigma'_k) = (a_k, \mu_k, \sigma_k) + (\delta_{a,k}, \delta_{\mu,k}, \delta_{\sigma,k})$, in such a way that $G(a'_k, \mu'_k, \sigma'_k)$ fits S_k better. The value $G(a'_k, \mu'_k, \sigma'_k)$ can be approximated by a linear function of $\delta_k = (\delta_{a,k}, \delta_{\mu,k}, \delta_{\sigma,k})^T$ as

$$G(a'_k, \mu'_k, \sigma'_k) \approx G(a_k, \mu_k, \sigma_k) + J(a_k, \mu_k, \sigma_k)\delta_k \tag{2}$$

where $J(a_k, \mu_k, \sigma_k) = \left[\frac{\partial G}{\partial a}\frac{\partial G}{\partial \mu}\frac{\partial G}{\partial \sigma}\right](a_k, \mu_k, \sigma_k)$ is the Jacobian of $G(a_k, \mu_k, \sigma_k)$ containing the derivatives of the Gaussian function with respect to its parameters:

$$\begin{aligned}
\left[\frac{\partial G}{\partial a}\right](f) &= e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}} \\
\left[\frac{\partial G}{\partial \mu}\right](f) &= \frac{a_k(f-\mu_k)}{\sigma_k^2} e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}} \\
\left[\frac{\partial G}{\partial \sigma}\right](f) &= \frac{a_k(f-\mu_k)^2}{\sigma_k^3} e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}}
\end{aligned} \tag{3}$$

The parameter δ_k that minimizes the error for the k-th Gaussian is given by

$$[J(a_k, \mu_k, \sigma_k)^T J(a_k, \mu_k, \sigma_k)]\delta_k = J(a_k, \mu_k, \sigma_k)^T [S_k - G(a_k, \mu_k, \sigma_k)] \tag{4}$$

where S_k is the k-th spectral sub-band and $J(a_k, \mu_k, \sigma_k)$ is the Jacobian of the k-th Gaussian $G(a_k, \mu_k, \sigma_k)$.

An important aspect of this algorithm is the stopping criterion in the search of a minimal solution, defined according to the rate of change of the absolute error $\mathcal{E}_k$ defined as

$$\mathcal{E}_k = ||S_k - G(a'_k, \mu'_k, \sigma'_k)||^2 \tag{5}$$

The estimation of each Gaussian ends when the variation of the error $\mathcal{E}_k$ between successive iterations is smaller than a given threshold, i.e. if $\Delta\mathcal{E}_k = |\mathcal{E}'_k - \mathcal{E}_k| \leq \epsilon$ is verified. The following algorithm estimates the parameters $(a_k^*, \mu_k^*, \sigma_k^*)$ of the k-th Gaussian that best fits the k-th spectral sub-band S_k. It depends on the threshold ϵ and on the initial values of the Gaussian parameters.

Algorithm 1.2 $(a_k^*, \mu_k^*, \sigma_k^*)$ = `Gaussian_Fitting(`S_k`,`(a^0, μ^0, σ^0)`,`ϵ`)`

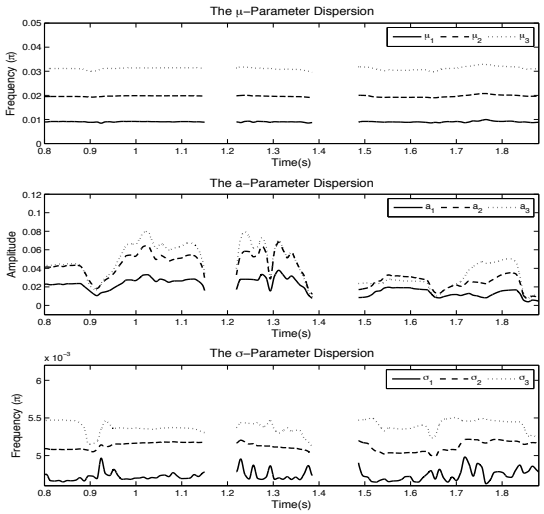


Fig. 1. *Dynamic Evolution of Gaussian Parameters.*

1 $\mathcal{E}_k^0 \leftarrow ||S_k - G(a_k^0, \mu_k^0, \sigma_k^0)||^2$
2 $\Delta\mathcal{E} \leftarrow \infty$
3 $(a_k, \mu_k, \sigma_k) \leftarrow (a_k^0, \mu_k^0, \sigma_k^0)$
4 **while** $\Delta\mathcal{E} > \epsilon$
5 Calculate $J(a_k, \mu_k, \sigma_k)$ using Eq. 3
6 Calculate δ_k by Eq. 4 using $J(a_k, \mu_k, \sigma_k)$
7 $(a_k, \mu_k, \sigma_k) \leftarrow (a_k, \mu_k, \sigma_k) + \delta_k$
8 Calculate $\mathcal{E}_k$ by Eq. 5
9 $\Delta\mathcal{E} \leftarrow |\mathcal{E}_k^0 - \mathcal{E}_k|$
10 $\mathcal{E}_k^0 \leftarrow \mathcal{E}_k$
11 **end while**
12 $(a_k^*, \mu_k^*, \sigma_k^*) \leftarrow (a_k, \mu_k, \sigma_k)$

Figure 1 displays the time evolution of the parameters $(a_k^*, \mu_k^*, \sigma_k^*)$ computed by Algorithm 1.1 on a sentence uttered by an arbitrary speaker. This evolution should be soft along the time, since the input sound is a vowel sustained /a/. In order to represent a set of L Gaussians, $3L$ parameters are used, which can become an expensive model depending on the application. The next section discusses alternative ways of representing the sum of Gaussians that keep some of these parameters fixed, making the size of the model smaller.

2.2 Parameter Constraints

An alternative use of sums of Gaussians to model spectral envelopes using fewer parameters consists of keeping some of the parameters μ and/or σ fixed as global

values, and to optimize the remaining parameters with respect to fitting the corresponding spectral sub-bands.

In this case, given one or two fixed parameters, we have to adjust the Jacobian matrix to maintain just the remaining (free) parameters within its structure. For instance, if we consider that μ is fixed ($\mu_k = c_k = \text{center}(B_k)$), then the reduced Jacobian used by this method is defined as

$$J(a_k, \sigma_k) = \left[\frac{\partial G}{\partial a}\frac{\partial G}{\partial \sigma}\right](a_k, \sigma_k) \tag{6}$$

Notice that the Jacobian matrix $J(a_k, \sigma_k)$ is a submatrix of the complete Jacobian defined in Equation 3, where the derivative with respect to μ has been eliminated. Other definitions have to be similarly adapted; for instance, in the above example $\delta_k = (\delta_{a,k}, \delta_{\sigma,k})^T$ and $\mathcal{E}_k = ||S_k - G(a'_k, \sigma'_k)||^2$. The optimal parameters will be given by a similarly modified version of Equation 4; considering μ fixed this modified equation will be:

$$[J(a_k, \sigma_k)^T J(a_k, \sigma_k)]\delta_k = J(a_k, \sigma_k)^T [S_k - G(a_k, \mu_k, \sigma_k)] \tag{7}$$

Everything else works exactly the same way. It should be easy to see how this method would work for σ fixed ($\sigma_k = \text{bandwidth}(B_k)$) and (a, μ) free.

A special case of parameter space reduction occurs when both μ and σ fixed. In this case the only free parameter would be a, and we would have $J(a_k) = \frac{\partial G}{\partial a}(a_k)$ and $\delta_k = \delta_{a,k}$ would have to verify

$$J(a_k)^T J(a_k)\delta_k = J(a_k)^T \left[S_k - G(a_k, \mu_k, \sigma_k)\right].$$

This special case where μ and σ are both fixed could also be addressed by a much more efficient method, where the amplitudes of all Gaussian components are estimated in one step, in order to minimize the overall error of the approximation $\|S - G(a, \mu, \sigma)\|$ where $G(a, \mu, \sigma) = \sum_{k=1}^{L} G(a_k, \mu_k, \sigma_k)$. If we maintain the restriction $a_k > 0$ for all k, this problem is reduced to a Non-Negative Least Squares Estimation - NNLSE [10].

Analogously, the same process of jointly estimating Gaussian amplitudes with μ and σ fixed can be used as a refinement step in each algorithm previously proposed. Obviously this implies in raising processing costs, as can be seen in Table 1. Here we present the Mean Square Error of each processed frame in an arbitrary signal, and also the number of iterations of the innermost loop, with and without the NNLSE refinement step. The conditions of this test will be described on Sec. 4 using the sawtooth synthetic signal.

It is seen in this table that all methods have a similar behavior, where the refinement step produces better solutions at the cost of increased computation.

3 Application on Pitch Shifting

In this section, we will discuss the application of the proposed models to pitch shift a voiced signal using the Harmonic plus Stochastic Model. A pitch shifting

Table 1. *Comparison of the proposed algorithms.*

Method	MSE		# iterations	
	w/out NNLSE	w/ NNLSE	w/out NNLSE	w/ NNLSE
Fixed-Band	0.0008	0.0006	22867	25341
Fixed-σ	0.0038	0.0017	23801	26275
Fixed-μ	0.0054	0.0017	21841	24315
Fixed-$\{\mu,\sigma\}$	0.0098	0.0022	7434	22266

algorithm modifies the fundamental frequency f_0 of a harmonic sound while preserving its spectral envelope. This application is convenient to understand perceptually the spectral decomposition in Gaussian sums and is used to assess the accuracy of the model.

3.1 Harmonic plus Stochastic Model - HSM

In this work, the system of Erro et al. [9] is used as baseline. This one is a popular model for analysis, synthesis and modification of speech is known as *Harmonic plus Stochastic Model – HSM*. This robust and compact system for speech signal representation splits a signal into static frames and decomposes them in harmonic and stochastic parts.

If a frame s is evaluated as non-harmonic, that is, *unvoiced*, only the stochastic part s_e is returned and the frequency f_0 is set to zero. The spectral envelope is modeled from the smoothed version of the spectrum $|S_e|$, using cubic splines interpolation of spectral peaks. These peaks are estimated by applying of a maximum filter with a lag corresponding to $60Hz$ in the discrete spectrum. Then, the sum of Gaussians $G_e(a, \mu,\sigma)$ is obtained from $|S_e|$.

If s is evaluated as harmonic, both harmonic and stochastic parts s_h and s_a are returned. $G_e(a, \mu,\sigma)$ is estimated from the stochastic part by the same method used in the non-harmonic case. For the harmonic part, the method returns a set of L amplitudes A_l and corresponding initial phases ϕ_l for each harmonic $l \in \{1, \ldots, L\}$, associated with a fundamental frequency f_0. We then produce a smoothed version of the harmonic spectrum $|S_h|$ by $15\times$ upsampling the amplitudes A_l using cubic spline interpolation. This smoothed harmonic spectrum is then modelled using a sum of L Gaussians, obtaining a model $G_h(a, \mu,\sigma)$ that will be directly used for modifying the pitch.

Suppose that the desired pitch shifting factor is ρ, i.e., the new fundamental is $f_0' = \rho f_0$ and all harmonic components of the modified signal will have frequencies $f_l' = \rho f_l,\ \forall l$. In order to preserve the spectral envelope (and consequently the formants of the voice), we obtain the magnitudes A_l using downsampling of the spectrum $|S_h|$. The modulated reconstruction of all harmonic amplitudes from the sum of Gaussians $G_h(a, \mu,\sigma)$ is immediate: A_l' will be defined as $A_l' = [G_h(a, \mu,\sigma)]\,(f_l')$. Since the stochastic information is supposed to remain unchanged in this operation, it is sufficient to change the values A_l and ϕ_l of each frame to modulate the pitch.

Although the spectral envelope is supposed to remain unchanged, some care must be taken to ensure the naturalness of the processed sound, and the phase update is crucial in this process. Ideally, given the k-th frame it is desired to define $f_l^{(k)}$ so that $\phi_l^{(k)} = \phi_l^{(k-1)} + \frac{2\pi N}{R} f_l^{(k)}$. Moreover, phase fluctuations existing between adjacent frames must be preserved to maintain the naturalness of sound. Consider that

$$\hat{\phi_l}^{(k)} = \phi_l^{(k-1)} + \frac{2\pi N}{R} l f_0^{(k-1)}.$$

Defining the phase fluctuation as $d_{\phi_l}^{(k)} = \phi_l^{(k)} - \hat{\phi_l}^{(k)}$, we can rewrite $\phi_l^{(k)}$ as function of $d_{\phi_l}^{(k)}$, so that

$$\phi_l^{(k)} = \phi_l^{(k-1)} + (d_{\phi_l}^{(k)} + \frac{2\pi N}{R} l f_0^{(k-1)}) = \phi_l^{(k-1)} + \frac{2\pi N}{R} f_l^{(k)}. \quad (8)$$

It is seen that $f_l^{(k)} = l f_0^{(k-1)} + d_{\phi_l}^{(k)} \frac{R}{2\pi N}$.

For pitch shifting, $f_l^{(k)}$ is multiplied by a modulation factor, and the current initial phase is reconstructed from the previous one (Eq. 8) preserving the phase difference of the k-th harmonic, $d_{\phi}^{(k)}$.

4 Evaluation

A preliminary experiment was made using 8 sentences in Spanish, each recorded by 4 male and 4 female speakers. The experiment was split into two parts, corresponding to objective and subjective evaluations.

4.1 Objective Evaluation

In this preliminary experiment we have calculated Mean Square Error (MSE) between given synthetic signals with known spectral envelopes and estimated spectral envelopes using several envelope estimation techniques. The input signal used is a sawtooth signal, whose spectral envelope has a $\frac{1}{f}$ overall shape that can be directly obtained using the continuous-time Fourier Transform. Table 2 below presents MSE values and average number of iterations for each method. The threshold parameter used in these methods was $\epsilon = 10^{-5}$. In this experiment, the amplitude refinement step using NNLSE presented in Sec. 2.2 was not used, except in the Fixed-[μ,σ] technique, which is precisely NNLSE applied to the amplitudes of the Gaussian components. Besides the four novel methods presented in this paper, we also compare the results with the spectral envelopes obtained with Cepstrum and LPC models of similar size. As we can see, the Fixed-Band method presented the lowest errors, although Fixed-[μ,σ] may be more useful with respect to computational time.

This objective experiment considers the overall fitness of the spectral envelope for each proposed model. However, it might be useful to consider specifically the fitness of these models with respect to the harmonic peaks of a voiced signal. For this reason, a simple perceptual experiment has been made in order to subjectively evaluate the naturalness of a modulated signal.

Table 2. *Avg. MSE and Avg. # of Iterations for each method.*

Method	Avg. MSE	Avg. # of Iterations
Fixed-Band	0.0025	20693
Fixed-σ	0.0033	21496
Fixed-μ	0.0061	19753
Fixed-[μ,σ]	0.0033	7340
CEPSTRUM(24)	0.0729	2384
LPC(24)	0.0318	2384

4.2 Subjective Evaluation

This subjective evaluation aims to analyse the naturalness of the modified voice signal after pitch modulation. The pitch shifting was chosen as an illustrative application due to the stronger link of pitch shifting with the spectral envelope.

The perceptual test consisted of pitch shifting 2 sentences uttered by 2 speakers by factors of 0.8 and 1.2. Table 3 presents the Mean Opinion Score (MOS) of 17 listeners for each method. Listeners were asked to rank the results of the different methods on a 5-point scale. The method contained in the baseline system of Erro et al. [9], using Discrete all-pole modeling (DAP) with 14 coefficients, that is a kind of LPC modeling using the autocorrelation of harmonic peaks; and the TD-PSOLA method, were also considered in this experiment.

Table 3. *MOS subjective evaluation.*

Method	Fem.(0.8)	Fem.(1.2)	Male(0.8)	Male(1.2)
Fixed-Band	4.44	3.69	2.94	3.69
Fixed-σ	3.94	3.88	2.75	3.69
Fixed-μ	3.69	3.38	2.56	3.25
Fixed-[μ,σ]	3.94	3.81	2.56	3.75
DAP	3.40	3.20	3.81	3.88
TD-PSOLA	2.88	3.13	4.06	3.63

As can be seen in Table 3, these methods achieve better results with signals that contains fewer harmonics (female voices), but MOS values are lower on signals with many harmonics (male voices). These lower values may be due to problems in the phases estimation of each frame in the reconstruction step, and will be better addressed in future works.

5 Conclusion

In this paper[3], a new parametric spectral estimation method was introduced assuming that the spectral envelope can be represented as a sum of Gaussians.

[3] This research was sponsored by CAPES and FAPESP grant 2008/08632-8.

Four models were presented for modeling a continuous approximation of the spectral envelope, which is controlled by amplitude, center frequency and bandwidth parameters, which can be defined using fixed values or optimized as free parameters using a modification of Marquardt's algorithm. Two experiments were presented to evaluate these models within an objective framework using synthetic signals with ground-truth spectral envelopes, and in a subjective scenario considering a pitch shifting application of real voice signals. The results obtained indicate that sums of Gaussians seems to be a promising spectral modeling tool for representing spectral envelopes of voice signals.

References

1. P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1229–1232.
2. B. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *Journal of Signal Processing*, 2007.
3. E. Godoy, O. Rosec, and T. Chonovel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1313–1323, 2012.
4. H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1303–1306.
5. P. Zolfaghari, S. Watanabe, A. Nakamura, and S. Katagiri, "Bayesian modelling of the speech spectrum using mixture of gaussians," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–553.
6. E. Godoy, O. Rosec, and T. Chonavel, "Speech spectral envelope estimation through explicit control of peak evolution in time," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 209–212.
7. A. Goshtasby and W. D. O'Neill, "Curve fitting by a sum of gaussians," *CVGIP: Graphical Model and Image Processing*, vol. 56, no. 4, pp. 281–288, 1994.
8. D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.
9. D. Erro, A. Moreno, and A. Bonafonte, "Flexible harmonic/stochastic speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, 2007.
10. D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

Towards a Robust Dynamic Frequency Warping Text-Independent Voice Conversion System

Tudor-Cătălin Zorilă[1], Daniel Erro[2], Yannis Stylianou[3,2], and Inma Hernáez[2]

[1]Telecommunication Department, Politehnica University of Bucharest, Romania
[2]Aholab Signal Processing Laboratory, University of the Basque Country, Bilbao, Spain
[3]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
ztudorc@gmail.com,{derro,inma}@aholab.ehu.es,yannis@csd.uoc.gr

Abstract. In this work we investigate several issues related to the use of Dynamic Frequency Warping in the context of text-independent voice conversion. For this type of systems, given an average spectral representation of each acoustic/phonetic class, dynamic programming is applied to compute the best alignment path between the frequency axis of the source and target speakers. In order to increase the robustness of the system, we suggest estimating such average spectral information using the Multi-Frame Analysis framework, while we compare several different local slope constraints for the dynamic programming procedure. Objective measurements show that the suggested approach provides better results than a state-of-the-art histogram-based solution in transforming the source spectral envelope towards the target one for all the dynamic programming constraints we considered.

Keywords: Voice conversion, voice quality, dynamic frequency warping

1 Introduction

In a society more and more interested in the latest technological breakthroughs, speech technologies are fast evolving to keep up with the trends. Telephony, music, film and computer games industries are insatiable when it comes to new speech technologies applications. In this context, voice conversion (VC) has emerged as a powerful technology that allows to modify the voice produced by one speaker (the source speaker) in such manner that it is perceived as that of another speaker (the target speaker), thus providing artificially generated speech with a high degree of flexibility. To achieve this goal, the speech individuality of the target speaker has to be modeled and transferred to the source speaker. The speech individuality involves both segmental and supra-segmental speech features, but there is a general agreement that the spectral envelope plays the determinant role among them [1-7]. In this paper the spectral envelope conversion will be referred to as 'voice conversion'.

In a typical VC system, mapping functions between source and target spectral envelopes are trained using speech utterances recorded from the source and target speakers. In the case of parallel training corpora, the speakers record the same collection of utterances, thus facilitating the search for a correspondence between source

and target features. In practice, it is not always possible to obtain parallel corpora from the involved speakers (in cross-lingual VC, for instance). As result, nonparallel training methods have been proposed to make VC systems text-independent [3].

Although probabilistic mapping methods based on Gaussian mixture models (GMMs) have prevailed over other VC methods [1-4], they have some important limitations, the most remarkable one being known as over-smoothing [4]. This phenomenon is produced by the restricted capability of a statistical VC function to capture the source-target correspondence accurately. Instead, the converted features yielded by the system tend to be too close to the mean values of the corresponding acoustic class. Consequently, the generated speech lacks naturalness and variability. Another source of quality loss which is inherent to GMM-based VC systems is the use of low dimensional spectral envelope parameterizations provided by vocoders.

Dynamic Frequency Warping-based VC method (DFW) [5, 8] is a non-parametric approach that maps significant points of the frequency axis of the source speaker into that of the target speaker, thus offering the premises of a higher perceptual quality due to the fact that spectral details are preserved during the transformation stage. In exchange, the individuality of the target speaker is not well transferred by pure DFW-based conversion approaches because the relative amplitude of the different spectral bands is not modified. In recently suggested DFW-based VC systems a suitable amplitude correction step is introduced after DFW to cope with this problem [6-7]. Furthermore, the solution proposed in [7] allows text-independent VC by training frequency warping functions from single representatives of the content of the phonetic classes. The training procedure of such a DFW-based system consists of three steps: (i) the training data are segmented into a number of acoustic/phonetic classes; (ii) for each class, using only the data therein, representative spectral information (RSI) is calculated for both the source and the target speaker; (iii) the frequency warping path that produces minimum-distortion between the two RSI is determined. The main disadvantage of this system is that, even when phonetic labels are available (as it is assumed in this work), the accuracy of the DFW-based VC system is strongly conditioned by the way steps (ii) and (iii) are implemented.

A high quality VC system implies estimating robust class-dependent RSI, which is related to robust spectral envelope estimation. In this work, we assume that a spectral envelope represents the magnitude of the frequency response of the filter which models speaker's vocal tract. In this context, for the voiced speech, a good/robust class-dependent RSI should be fundamental frequency independent. This leads to the main problem to be solved, namely how to combine the contribution of multiple training frames into a single RSI. One possible solution is to average data sets [5], but this approach will lead to flat spectrum, so spectral details will be lost during transformation. Godoy et al. [7] suggested a better way to solve the problem by calculating class-dependent frequency histograms from meaningful (formant related) peaks of the magnitude spectra. These histograms are used as class-dependent RSI, where the k-th element is the probability that a formant peak is found in the k-th frequency bin. However, the main disadvantage of this approach is its sensitivity to the analysis method used to construct the histograms. The solution proposed by Godoy et al. will be denoted as Hist-VC from now on. In this paper we suggest a more

robust way of estimating the RSI of each class, i.e. Multi-Frame Analysis (MFA) [9]. Throughout the paper this method is denoted MFA-VC.

Given the source and target RSI that corresponds to a specific class, frequency alignment is carried out by means of an automatic dynamic programming technique. Due to vocal tract length and shape differences, aligning the speakers' RSI is not a trivial task (even manually!). Therefore, in this work several local slope constraints in the implementation of the DFW are investigated.

The paper is organized as follows. Section 2 provides a brief description for the methods used throughout the paper. Section 3 describes the experimental setup and presents the results. Finally, section 4 summarizes the conclusions and future work.

2 Methods Towards DFW-based VC

Figure 1 presents the training stage block diagram for the suggested MFA-VC system (solid line). A baseline approach for a DFW-based VC, Hist-VC (similar to the one described in [7]), is also shown (dashed line) for comparison purposes.

Segments of speech signals from the same phonetic/acoustic class (for both speakers) are divided into frames of length twice the pitch period and then the magnitude spectrum of these frames is computed using DFT. Next, a first Peak Picking is applied on the DFT magnitude spectra. The spectral peaks are pairs of magnitude and frequency bins, which correspond to the harmonics in voiced frames. These peaks are used as input for both MFA-VC and Hist-VC systems. The MFA technique is briefly presented in the following sub-section.

The block called "Spline + Peak Picking + Hist" is the module that generates the RSI for the Hist-VC system. It uses the peaks provided by the Peak Picking block and it computes spectral envelope by spline interpolating between these peaks (in voiced frames, this eliminates the pitch dependencies). Then, a second peak-picking is used to extract the local maxima from the estimated spectral envelope, thus the frequencies of these new peaks are an estimation of the formants' location. After analyzing all the training frames, for each phonetic class, histograms are built to model the distribution of the formants in frequency.

The DFW block aligns the RSI for both systems and generates an optimal frequency warping path per phonetic/acoustic class. The Amplitude Correction (AC) compensates for the remaining differences between the frequency-warped source and target spectral envelopes [7].

Using histograms as RSI is advantageous with respect to an average magnitude spectra solution because it removes the effect of spectral tilt and avoids spectral flattening. However, it is sensitive to the analysis parameters, such as: the analysis window type and length, the masking functions for selecting the representative peaks from the spectrum or the smoothing approach used to obtain the histograms.

As alternative, the suggested MFA-based VC uses the powerful Multi-frame Analysis core to extract robust RSI and offers in the same time the ability to generate robust AC functions (see Fig. 1). The dual gain from using MFA (in both estimation

robustness and system complexity reduction) can be boosted by the usage of a proper (physically motivated) DFW technique.

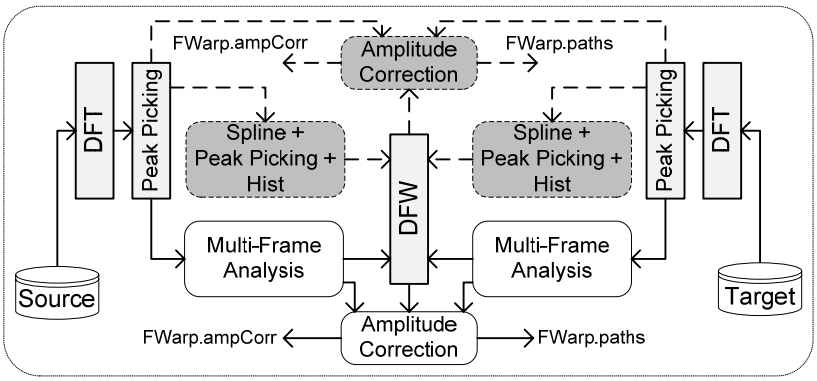


Fig. 1. Training stage block diagram for the suggested MFA-VC (solid line) and a Hist-VC systems (dashed line)

The following sections present the MFA spectral envelope estimation, the DFW and the AC internals.

2.1 Multi-frame Analysis Spectral Envelope Estimation

Due to quasi-periodic nature of voiced speech and due to changes of the fundamental frequency during speaking, the simple averaging of the magnitude spectrum from each acoustical class does not offer a robust estimator for the RSI of an acoustic/phonetic class. Assuming that such information exist (or it is valid) for a class, we may consider that a specific frame provides simply a sample of that information. However, the usage of a single frame will not provide a good estimation of the average spectral representation of an acoustic class. This is actually similar to the problem of computing the mean of a (multidimensional, in this case) stochastic variable using only one realization. Multi Frame Analysis offers a mechanism to estimate in a robust way this average spectral information (avoiding however any smoothing and whitening that a simple average operator will produce) [9].

MFA has been presented before in the context of speech synthesis but not in the context of voice conversion.

MFA uses cepstrum and Least Squares Estimation for all the frames from a phonetic category. Denoting $c[n]$ as the cepstrum for a discrete sequence $x[n]$, then the log magnitude spectral envelope can be approximated as

$$\ln|X(f)| = \sum_{n=-p}^{p} c[n]\cos(2\pi fn) \tag{1}$$

where p is the cepstrum's order.

Supposing M frames from a specific acoustic/phonetic class, these will be prior analyzed in terms of their log-amplitudes a_k^l and frequencies f_k^l (in this paper we consider the spectral peaks extracted by the Peak Picking block – see Fig.1), where the upper index denotes the l-th spectral peak and the lower index denotes the k-th speech frame. The cepstrum coefficients are estimated using least squares, considering the error ε to be minimized as:

$$\varepsilon = \sum_{k=1}^{M} \sum_{l=1}^{N_k} \frac{w(f_k^l)}{N_k} \left[a_k^l - d_k - \sum_{n=-p}^{p} c[n] \cos(2\pi n f_k^l / F_s) \right] \quad (2)$$

where $w(f)$ represents a weighting function used to put more emphasis on the lower part of spectrum, N_k is the number of harmonics in each frame, d_k is an offset factor used to correct the total power of each frame such that ε gets minimized and F_s is the sampling frequency.

In (2), the minimization of ε is a non-linear procedure since the error depends on $c[n]$ and d_k. Assuming an initial value for d_k, an iterative approach is adopted [9]. The iterations stop when the error is lower than a predefined threshold, providing the MFA-based p order cepstrum (1).

2.2 Dynamic Frequency Warping

To automate the process of finding the correspondence between RSI, dynamic programming technique is used [5, 7-8, 10-11]. The correspondence between RSI is a core component of a DFW-based VC system, as result we are considering different ways of implementing the dynamic programming in order to achieve the best physically meaningful warping paths. As result, in this paper three variations for the dynamic programming are considered, with different complexity and flexibility.

The first DFW technique accounted (denoted as DFW_1) has the maximum degree of freedom in searching for the optimal warping paths. Its local slope constraint is given by:

$$C_1[i,j] = \min \begin{Bmatrix} C_1[i-1,j-1] \\ C_1[i-1,j] \\ C_1[i,j-1] \end{Bmatrix} + S[i,j] \quad (3)$$

where C_1 is the cost matrix associated with the dynamic programming and S is the similarity matrix between the two RSI. For the similarity matrix the following distance is used (for all the methods):

$$S[i,j] = \left| RSI_1[i] - RSI_2[j] \right|,\ i = \overline{1, N_1},\ j = \overline{1, N_2} \quad (4)$$

where N_1 and N_2 are the lengths for RSI_1 and RSI_2, respectively. In this context, the goal of DFW is to find the minimum cost warping path $P = \{(i_1, j_1),(i_2, j_2),\ldots,(i_K, j_K)\}$ [8, 10-11].

The DFW_1 freedom for the warping path search is not always an advantage because it could lead to physically unsustainable warping functions. Therefore, quite complex global slope constraint policies are considered very often [10], with impact on the overall performance of the system. An intermediate step towards such a complex approach is by using a "built-in" global slope constraint (by the way the local slope is defined), with limited impact on the system's performance. Our extended experiments using different well known speech recognition local slope constraints scenarios [10-11] recommend the following form (referred to as DFW_2):

$$C_2[i, j] = \min\begin{Bmatrix} C_2[i-1, j-1] \\ C_2[i-2, j-1] \\ C_2[i, j-1] \end{Bmatrix} + S[i, j] \tag{5}$$

The third tested technique (denoted as DFW_3) is based on DFW_1, but introduces a higher degree of control to select the minimum cost frequency warping paths, following the work of Matsumoto and Wakita for non-uniform talker normalization [8]. By imposing a complex set of restrictions (boundary, continuity and nonlinearity conditions) the frequency warping functions are searched within a physically significant area. As result, with a higher computational price, physical meaningful warping paths can be obtained between RSI. The boundary conditions used in our implementation are 0.6 and respectively 1.4 for the slopes defining the search area. The values are chosen to cover a large dynamic for the ratios between speakers' min and max vocal tract lengths [8]. The advantage of this method is that speaker adaptive estimates for the vocal tract lengths could be easily integrated into the framework, therefore enhancing the DFW and promising higher VC quality.

2.3 Amplitude Correction

Amplitude correction is implemented as a corrective filter (CF) [12]. For the MFA-VC system the AC is implemented as a 12 order CF, computed from the difference between the frequency-warped source and target average spectral envelopes. The average spectral envelopes are generated from the MFA estimates. In the case of Hist-VC, an approach closer to the implementation mentioned in [7] is employed. The same 12 order CF is used, but in this case the average spectral envelopes (frequency warped for the source) are computed by means of discrete cepstrum [7, 13].

3 Evaluation and Results

In this experiment we are using CMU_ARCTIC speech database [14], consisting of phonetically balanced US English single speaker recordings at 16 kHz sampling frequency. For the purpose of spectral envelope conversion we have selected four speakers: two female speakers (*slt* and *clb*) and two male speakers (*bdl* and *rms*). All the voice conversion directions are considered: male to female (bdl to slt, denoted *bdl2slt*), female to male (clb to rms, denoted *clb2rms*), male to male (rms to bdl, denoted *rms2bdl*) and female to female (slt to clb, denoted *slt2clb*). A number of 100 parallel training sentences per speaker are selected, while a different set of 20 randomly selected sentences are considered for testing purposes.

During the training stage, the speech frames are extracted from segments defined by the labels provided by the ARCTIC's phonetic segmentation. The segmentation consists of 44 American English phonetic classes. The voiced speech segments are analyzed using a Hanning window with length twice the average pitch period, while for unvoiced segments a fixed 100 Hz pitch frequency is imposed. The training is done using only frames from the middle section of the segments to avoid boundary artifacts.

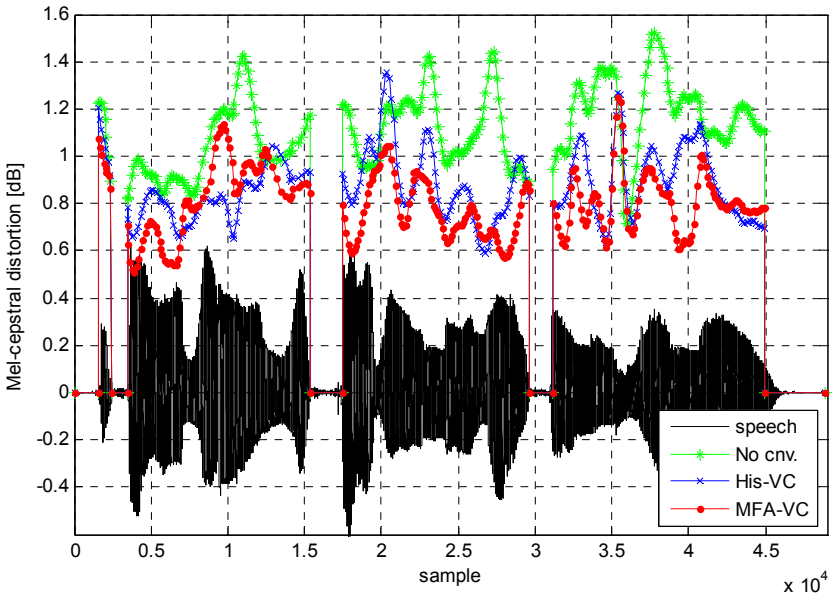


Fig. 2. The MCD scores per voiced frames for the "If I ever needed a fighter in my life I need one now" sentence, using DFW2 approach and the following methods for the *bdl2slt* VC direction: no conversion applied (the green with stars curve), Hist-VC (the blue with x-s curve) and MFA-VC (the red with dots curve)

The MFA-VC system is objectively compared against the Hist-VC for different speaker pairs and for different dynamic programming methods. The objective measure used is the average Mel-cepstral distortion (MCD) between the converted and the target spectral envelopes [4]:

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{K}\left(mc_k^{tgt} - mc_k^{src\to tgt}\right)^2} \quad (6)$$

The average MCD is computed only for voiced frames.

In Fig.2 is presented an example of spectral envelope conversion between a male and a female speaker in terms of MCD over time, using MFA-VC and Hist-VC systems. For comparison purposes the "no conversion" MCD is also depicted. It is obvious that the voice conversion is effective because the MCD drops. The general trend shows that MFA-VC offers lower distortions than Hist-VC. Another observation is that the distortion contours present an oscillatory behavior, with minimum corresponding in general to the center of the phonetic segments. This is normal because in this implementation the transformation applied per acoustic class is the same for all the frames inside a segment, thus the error is higher during segment transitions. Inter-segment interpolation between warping paths and CFs will be considered in our future work.

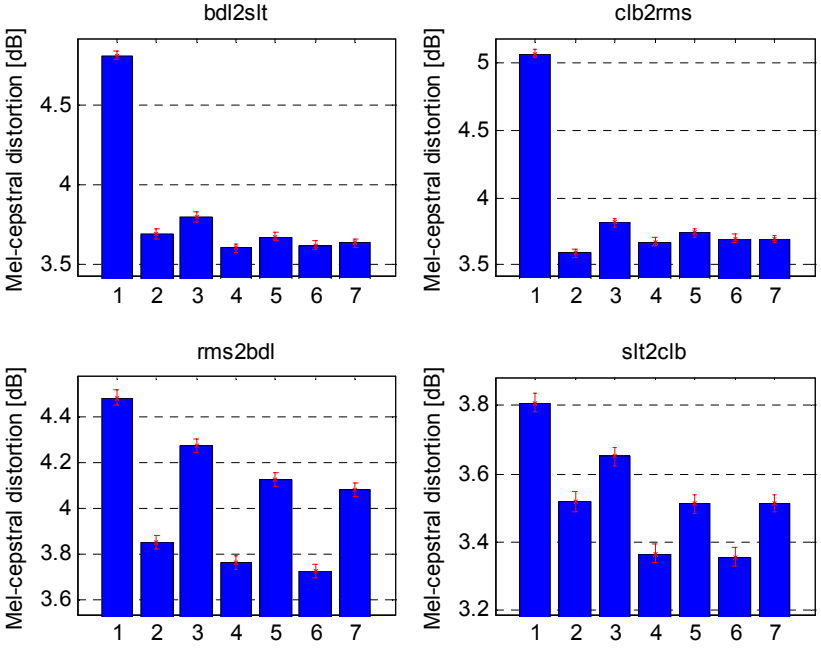


Fig. 3. Average MCD scores with 95% confidence intervals for all the VC directions and for different DFW approaches, in the case of MFA and histogram based VC systems, as follows: 1– no conversion, 2 - MFA-DFW1, 3 – Hist-DFW1, 4 – MFA-DFW2, 5 – Hist-DFW2, 6 – MFA-DFW3, 7 – Hist-DFW3

Table 1. Objective comparison between the MFA-VC and Hist-VC systems for the considered DFW techniques. The MCD values are averaged for all the VC directions

	No cnv.	MFA-DFW1	Hist-DFW1	MFA-DFW2	Hist-DFW2	MFA-DFW3	Hist-DFW3
Average MCD (dB)	4.54± 0.03	3.66 ± 0.03	3.88 ± 0.03	3.60 ± 0.03	3.76 ± 0.03	3.60 ± 0.03	3.73 ± 0.03

In Fig.3 the average MCD scores for all the VC directions are shown, with 95% confidence intervals values around ±0.03 dB. The lowest MCD for all the directions is obtained for the MFA-VC system. Furthermore, the superiority of MFA is maintained for all the considered DFW methods. As it is easy to see from Table 1, DFW_2 and DFW_3 offer the best performances. Taking into account the higher arithmetical complexity given by DFW_3, the optimal solution remains DFW_2. Even so, we believe that the power (in terms of robust and meaningful warping paths) given by DFW_3 is not completely unleashed. In our future work we will consider speaker adaptive parameters for the warping functions search area to increase the VC robustness and quality.

4 Conclusions and future work

Objective measures indicate that the suggested MFA-VC framework outperforms the Hist-VC opponent, both in terms of adopted warping strategy and conversion direction. By integrating prosody modifications as well, extended subjective tests between our framework and other state of the art VC systems are scheduled as a future work.

Acknowledgements. This work has been partially supported by the Romanian Ministry of Labor, Family and Social Protection (financial agreement POSDRU/88/1.5/S/61178), the Spanish Ministry of Science and Innovation (Buceador, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262; ZURE_TTS, SPE11UN081). Y.S. holds a Visiting Professorship from UPV/EHU.

References

1. Y. Stylianou, O. Cappé, E. Moulines, “Continuous probabilistic transform for voice conversion”, IEEE Trans. Speech and Audio Process., vol. 6, pp. 131-142, 1998
2. A. Kain, “High resolution voice transformation”, Ph.D. thesis, Oregon Health & Science University, 2001
3. A. Mouchtaris, , J. Van der Spiegel, P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 952-963, 2006

4. T. Toda, A.W. Black, K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory”, IEEE Trans. Audio, Speech, Lang. Process., vol. 15(8), pp. 2222-2235, 2007
5. H. Valbret, E. Moulines, J.P. Tubach, “Voice transformation using PSOLA technique”, Proc. ICASSP, pp. 145-148, 1992
6. D. Erro, A. Moreno, A. Bonafonte, “Voice conversion based on weighted frequency warping”, IEEE Trans. Audio, Speech, Lang. Process., vol. 18(5), pp. 922-931, 2010
7. E. Godoy, O. Rosec, T. Chonovel, "Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, pp. 1313-1323, 2011
8. H. Matsumoto, H. Wakita, “Frequency warping for nonuniform talker normalization”, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1979
9. Y. Shiga, S. King, "Estimating the Spectral Envelope of Voiced Speech Using Multi-Frame Analysis," in Proc. EUROSPEECH 2003, 2003, pp. 1737-1740
10. H. Sakoe, S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): p. 43-49
11. L. Rabiner, B.H. Juang, Fundamentals of speech recognition: Prentice Hall, 1993
12. Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," PhD Thesis, Ecole Nationale Supérieure des Télécommunications, 1996
13. O. Cappé, J. Laroche, E. Moulines, “Regularized estimation of cepstrum envelope from discrete frequency points”, IEEE Workshop on Apps. Signal Process. to Audio & Acoustics, pp.213-216, 1995
14. CMU ARCTIC speech synthesis databases. Online: http://festvox.org/cmu_arctic/

Poster Session P2

Rule-Based Pronunciation Modeling for Resource-Scarce Varieties

Simone Ashby, José Pedro Ferreira, Sílvia Barbosa, Catarina Silva

[1]Instituto de Linguística Teórica e Computacional (ILTEC), Lisbon, Portugal
{simone, zpferreira, silvia, catarina}@iltec.pt

Abstract. Narrowing the speech technology gap within and between nations requires, among other things, a strategy for coping with scarce linguistic resources. In this paper, we describe the ongoing research activities of the LUPo project to model non-standard varieties of spoken Portuguese from Mozambique and other resource-scarce countries and regions. We demonstrate how these models have been combined to form a rule-based architecture for generating accent-specific pronunciations, adapted from the Unisyn Lexicon for English. We describe some of the challenges of extending the Unisyn model to Portuguese, and of establishing models for non-standard varieties that receive considerable substratal interference through contact with local languages. Finally, we discuss some preliminary results, quality control efforts, and the merits and pitfalls of applying the Unisyn methodology to generate pronunciation dictionaries for resource-scarce varieties.

Keywords: rule based pronunciation modeling, pronunciation dictionaries, Portuguese, regional accents, Mozambique, resource-scarce

1 Introduction

Portuguese is the world's sixth largest language, with major concentrations of speakers in Angola, Brazil, Cape Verde, East Timor, Guinea Bissau, Mozambique, Portugal, and São Tomé and Príncipe. The LUPo (Portuguese Unisyn Lexicon) project is dedicated to: (1) modeling standard and non-standard varieties of spoken Portuguese from around the globe; (2) delivering a free, open-source tool for the automatic generation of accent-specific pronunciation lexica within the existing *Portal da Língua Portuguesa* (http://www.portaldalinguaportuguesa.org); and (3) providing the research community with a free, online, searchable databank dedicated to the description of regional varieties of spoken Portuguese.

While numerous resources exist for the dominant Brazilian (BP) and European (EP) Portuguese standards – including electronic dictionaries, phonetically aligned corpora, pronunciation lexica, and detailed published studies – less dominant varieties of Portuguese typically have far fewer linguistic resources at their disposal. This is especially true of Mozambique and other non-western Lusophone nations, despite the fact that L1 (first language) speakers are fast on the rise in such countries, where "Portuguese is viewed as the language of science, knowledge, and power", and that

which holds the most promise for obtaining employment and enhancing one's upward mobility [1].

With few available linguistic resources, the less dominant varieties see little in the way of language technology adaptation. Speech applications, which rely on the existence of good quality pronunciation dictionaries, continue to be designed with end-users from Lisbon/Coimbra, Rio de Janeiro, or São Paulo in mind. As a result, researchers active in promoting spoken language technologies for development in areas such as northeastern Brazil or the Mozambican capital of Maputo must find alternate means of establishing pronunciation dictionaries for these and other resource-scarce varieties of Portuguese.

The Unisyn model described in [2] offers a unique approach to modeling pronunciation variation through investment in an accent-independent lexicon and rule system for generating high quality accent-specific phonetic output. Within this model, the accent-independent or 'master' lexicon contains orthographic entries and their underspecified "phonemic" representations, encoded via computer-readable key symbols (see Figure 1). These can best be understood as 'metaphonemes', and take from the ideas in [3] as a means of "[a]bstracting away from phonetics [so] that a single lexicon can represent numerous different accents" [4]. Additional symbols are used for marking stress, syllable, and morpheme boundaries, as well as for linking phonemes to their graphical counterparts.

937	gorra	substantivo	{ g #o . r_rr == a }
102789	gorro	substantivo	{ g #o . r_rr == u_o }
104152	gosma	substantivo	{ g #O_o z_s . m a }
78696	gosmar	verbo	{ g o z_s . m #a } > r >
56651	gospel	substantivo	{ g O_o s . p #E_e 5_l }
6187	gostar	verbo	{ g o s . t #a } > r >
58007	gostinho	substantivo	{ g o s . t == #i . J_nh u_o }
32523	gosto	substantivo	{ g #o s . t == u_o }

Fig. 1. Sample master lexicon entries encoded in X-SAMPA, as adapted by LUPo. The numeric field signifies corresponding entries in the Portal da Língua Portuguesa and their record IDs. Syllable boundaries are marked with a dot '.', primary stress is denoted with a hash symbol '#', and orthographic linking is denoted with an underscore symbol '_'. The remaining non-alphabetic symbols denote morpheme boundaries.

Phonetic conversions involve the application of hand-written rules to the underspecified master lexicon representations, a system that is governed by a geographically organized regional accent hierarchy. In simple terms, this consists of a system of nodes corresponding to the levels COUNTRY, REGION, TOWN, and PERSON, for identifying which rules correspond to one or more of the varieties modeled. What is interesting about this hierarchy is the inheritance by each node of features from the higher level node, provided the inheritance is not broken by the introduction of competing rule(s) at a lower level. As such, the Unisyn framework enables the sharing of rules across geopolitical varieties, allowing for the representation of non-systematic relationships between branch nodes (e.g. the

realization of word-final /l/ as [ɬ] by a small sub-population of São Paulo speakers and among a much larger population of speakers in Angola, Cape Verde, Mozambique, Portugal, etc.), while gaining its economy from the broader notion that phonetic similarity is, to some extent, tied to geographical proximity (e.g. the widespread use of word-final [w] for /l/ in Brazil). It is also possible to represent sociolectal and idiolectal varieties by introducing rules at the low-level nodes that combine with a set of more generally applied rules to capture the unique segmental characteristics of a neighborhood, or an individual speaker.

The Unisyn framework not only offers a robust method for generating large-scale pronunciation dictionaries for multiple varieties, and handling out-of-vocabulary items, it is also extremely well suited for representing varieties with limited linguistic resources. Once established, the sets of hand-written, accent-specific rules and their lexical exceptions can be used to generate high-quality pronunciation dictionaries for use in ASR and TTS applications, and other spoken language technologies.

In the sections that follow, we describe the rule creation processes employed by LUPo for modeling the variety of Portuguese spoken in Maputo, Mozambique. We then show how these rules have been integrated into LUPo's pronunciation generator, whereupon we highlight some of the challenges of modeling non-standard varieties that receive considerable interference from local languages, and, more generally, of adapting the Unisyn model to Portuguese. A description of LUPo's preliminary results and evaluation criteria follows, along with an assessment of the Unisyn framework's suitability for creating pronunciation dictionaries for resource-scarce varieties.

2 Modeling the Maputo Variety

2.1 Data Collection

Given LUPo's aims and the fact that Portuguese is widely spoken throughout Mozambique's larger urban centers, inclusion of the Maputo variety (P_{Map}) was identified as one of the project's top priorities. As with other resource-scarce varieties, our initial efforts were spent collecting a corpus of read and spontaneous speech from male and female informants, aged 18-35, from the target area. The majority of speech recordings were conducted at the Eduardo Mondlane University in Maputo, with additional recordings performed at ILTEC, in Lisbon, Portugal. In total, we collected more than 21 hours of spoken data from 14 informants, nine whom regard Portuguese as their first language,[1] while the remaining five informants reported learning Portuguese during the early years of their primary education.

Materials for the elicitation of read speech are based on those in [5], along with an additional set of words and phrases deemed necessary for capturing other relevant

[1] In the cities of Mozambique, it is increasingly common to raise children with Portuguese as the L1. According to [6], Portuguese is spoken as a lingua franca by 50.37% of the population; an additional 10.7% regard Portuguese as their native language.

contexts. For the read speech elicitation task, informants were asked to read the individual phrases and sentences projected in front of them on PowerPoint slides. The elicitation of spontaneous speech data was conducted in the form of an oral questionnaire for obtaining general speaker information and attitudinal data. Recordings were performed with a Marantz digital voice recorder, with a microphone positioned on the table in front of the informant.

2.2 Accent Modeling and Rule Creation

By accent, we mean those features characteristic of a speaker's pronunciation. While future plans involve the modeling of suprasegmental characteristics, our current efforts are focused on describing the broad segmental features of P_{Map}. In terms of style, this model reflects a slower and more carefully articulated variety than that found in conversational speech.

Corpus-based accent models were developed at the segmental level by first establishing a list of contexts in which the metaphonemes described in Section 1 were likely to have a different surface realization for the variety being modeled. This was accomplished by listening to read-speech data for three of the L1 informants, creating a new list entry when alternate surface realizations were heard, and noting the observed realization(s) in the form of rules. The list of contexts was then expanded and modified using observations from data for the remaining informants, with a focus on those contexts which yielded variable output in the original three informant recordings analyzed. In cases where a metaphonemic context resulted in variable output, frequencies were calculated across informants, with an indication of the most stable variant per context.

For example, in words such as *Evangelho* 'gospel' and *melhor* 'better', the palatal lateral approximant /ʎ/ was realized by our Maputo informants as [j$l] or [ɬ$j] (with '$' denoting a syllable boundary) . For a very small number of speakers, /ʎ/ was additionally realized as the simplified [l], although not systematically. Being most common, [j$l] was selected as a variant of [ʎ][2], and the following rule was formed (here and throughout this paper using X-SAMPA notation): **do_L_jl_resyllabify**. This rule instructs LUPo's pronunciation generator to perform a conversion when a word in the master lexicon is encountered containing /ʎ/ (or 'L'), thus producing the resyllabified output forms 'Evang[ej$l]o' and 'm[ej$l]or'. This contrasts with output for standard BP and EP varieties, in which the /ʎ/ is preserved as a syllable onset.

Together, these rules – along with the features encoded in the master lexicon that are resistant to change, and the particular lexical exceptions that apply to a given variety – form a segmental accent model. In many cases, a rule already exists for one or more of the varieties modeled, in which case the applicability 'score' for that rule must be modified to include the new variety. Naturally, it is also of extreme importance that the rules be ordered. These and other topics related to LUPo's implementation of the Unisyn system are described in Section 3.

[2] *All* variants are organized per context and stored in LUPo's free, searchable, online databank (forthcoming) for potential use in linguistic studies and ASR applications.

Substratal Interference. The diglossia that defines Mozambique linguistically means that Mozambican Portuguese speakers are also fluent in one or more Southern Bantu languages. This is no less true of our P_{Map} informant population, which reported varying degrees of fluency in Changana, Chope, Gitonga, Ronga, Tshwa, and Zula, as well as French and English. The effects of local Bantu substrates on informants' spoken Portuguese varieties are often attested in the literature on Bantu linguistics, and can be modeled in the form of rules. E.g. do_d_dd_resyllabify: gemination of the word-final onset /d/ in penultimately stressed words, such as esperança[d$du] *esperançado* 'hopeful' and ansieda[d$di] *ansiedade* 'anxiety'; do_monothong: the monothongization of diphthongs in words such as r[u]nião *reunião* 'meeting' and ap[o$j]o *apoio* 'support'; and the combination do_devoicedC and do_V_aspirant_resyllabify for the realization of aspirated, devoiced consonants and vowel apocope in unstressed, word-final contexts, such as be[p^{h}] *bebo* 'I drink', and objeti[f^{h}] *objetivo* 'objective'.

Other phenomena are more complex. For example, like BP varieties, consonant sequences are predictably broken up by epenthetic vowels. The problem for P_{Map} lies in adapting the Unisyn's search and replace style regular expressions to properly identify these vowels, which assimilate with the vowel in a neighboring syllable. E.g. **o[mo$n]isciente** *omnisciente* 'omniscient', **interre[ge$n]o** *interregno* 'interruption', and **eru[pi$s]ão** *erupção* 'eruption'. A separate long-distance phenomenon occurs, in which the second in a non-contiguous series of voiced stops undergoes voice or place dissimilation. E.g. **obe[t]iência** *obediência* 'obedience', and **de[ð]ozinho** *dedozinho* 'little finger'. Contexts requiring look-ahead and look-behind methods of computation are described in Section 3.

As with a large number of Portuguese varieties, the rule for palatalizing the coda /s/ (i.e. **do_s_S**) takes effect before another consonant. However, there exists in P_{Map} a circumstance in which the deletion of a subsequent syllable causes the resultant [ʃ] to become realized as a syllable onset, or in which [ʃ] undergoes metathesis in the context of a neighboring sibilant. E.g. **inve[$ʃ*]igações** *investigações* 'investigations' (instead of **inve[ʃ$t]igações**), and **omni[s$ʃ]iente** *omnisciente* 'omniscient' (instead of **omni[ʃ$s]iente**). In these and other cases, LUPo's rule system must be carefully ordered and specified so that the necessary context is in place for applying the rule **do_s_S**, and so that predictable contexts for elision – e.g. /t/ in the dispreferred sequence [ʃt] – undergo subsequent resyllabification.

Variability. Inter- and intra-speaker variability issues must be accounted for in the treatment of any regional variety. However, given the degree of multilingualism described in Section 2.2.1, and the fact that the Portuguese language is still in the process of becoming indigenized in Mozambique, phonetic variability can be described as a defining feature of P_{Map}, thus requiring adequate representation in LUPo's pronunciation output.

Some of the contexts in which P_{Map} is particularly prone to yielding variable output include: realization of word-final /r/ as an alveolar trill [r], a flap [ɾ], a voiceless velar fricative [x], or as an elided segment [*]; non-systematic pronunciation of "silent" consonants /p/ and /k/ in words such as **ado[p]ção** *adopção* 'adoption', and **da[k]tilografia** *dactilografia* 'typing', which are no longer phonetically represented in BP and EP; realization of /t/ as a stop [t] or a homorganic fricative [θ]; realization

of the stressed nasal vowel /ẽ/ as a diphthong [ẽj] or a raised segment [ĩ]; and variable realizations of the open and close vowel pairs /ɛ, e/ and /ɔ, o/.

Implications for the Master Lexicon. As indicated, the metaforms that make up the master lexicon are abstract so as to minimize the number of rules necessary to generate target output for multiple regional varieties. Given the strong correspondence between phonemes and graphemes in Portuguese, the specification of segments was accomplished in a rather straightforward manner. For example, the metaform **{ a . b #O_o . b o . r a }** maps to the word *abóbora* 'pumpkin'.

However, master lexicon development also involves a level of decision making that is informed by the specific varieties being treated. Again, this can be handled in a relatively straight-forward manner if one already has a sufficient understanding of the regional variety in question. But, in the case of resource-scarce varieties, such as P_{Map}, not all of the important considerations for master lexicon development will be obvious until partial modeling is done for this and the other varieties concerned. For example, based on what we already know about the strong tendency for gliding in standard BP and EP varieties in words such as **en[sju]mar** *enciumar* 'to make jealous' and **contí[nwu]** *contínuo* 'continuous', it is tempting to encode these and other common types of gliding in the master lexicon to minimize the number of rules needed. With some modeling of P_{Map}, however, the need to preserve the original vowel sequences became clear so as to generate the appropriate output forms **en[si$u]mar** and **contín[nu*]**. Orthographic linking also emerged as particularly important in preserving the context for fully specified, word-final, nasal consonants, which occurs in P_{Map} among non-high-frequency words. E.g. the specification of **{ m #o . == e ~_m }** (instead of **{ m #o . == e ~ }**) for *moem* 'they grind' for yielding the output nasal vowel [ẽ] + [m#] by P_{Map} speakers, as opposed to just [ẽ#]. This type of encoding is essential for identifying the contexts of other orthographically motivated pronunciations, such as occurs with the "silent" graphemes <c, p> among P_{Map} speakers (*see also* Section 2.2.2).

3 Integrating P_{Map} Rules into LUPo

As indicated above, LUPo's rules take effect through application to the underspecified *metaforms* stored in the master lexicon, each of which corresponds to a lemma in the Portal da Língua Portuguesa (hereafter referred to as the *Portal*). A component file system is used to: define LUPo's regional accent hierarchy, assign rule scores for determining which rules apply to which varieties and the order in which they take place, define classes of sounds, govern phonotactic sequences, deal with lexical exceptions, and access word frequencies.

Integrating P_{Map} rules into LUPo first required ensuring that the rule did not already exist. P_{Map} shares a number of its rules with standard BP and EP, such as the aforementioned palatalization rule **do_s_S**, which characterizes the standard Rio de Janeiro and Lisbon/Coimbra varieties, and many other nonstandard varieties. The fact that LUPo already features rules for BP and EP meant that by simply switching these rules 'on' in the rule scores file, a significant part of the work was already done for

P_{Map}. By the same token, introducing new rules for P_{Map} facilitates the work of adding more regional varieties from Mozambique and other Luso-African countries that share a common store of allophones. This is very meaningful for adapting rule systems to resource-scarce varieties and establishing tailor-made pronunciation dictionaries in a relatively low-cost manner.

Still, the introduction of novel rules is not trivial, in part given the need for careful attention to rule ordering and resyllabification phenomena. These matters being carefully attended to in the original Unisyn lexicon, the model holds up well to Portuguese adaptation. Specifically, each rule has a *$ruleorder* assignment in the rule scores file that tells the system which rules apply to which varieties and in what order they should take effect. In terms of resyllabification, the Unisyn lexicon's search and replace method and system for marking syllable and morpheme boundaries are well suited for generating resyllabified output. Returning to the example *investigações*, from Section 2.2.1, the *$rulescore* for the newly created rule **simplify_cluster** was adjusted to take effect *after* the existing rule **do_s_S**, and override the syllable boundary specified in the master lexicon with a new one, yielding the output **inve[$ʃ*]igações**. As such, the Unisyn framework has also proved highly suitable for dealing with sandhi phenomena.

What is more challenging in terms of computation, logistics, and quality control is: (1) the need to account for complex substratal interference (see Section 2.2.1); (2) the important role spoken varieties play in informing the design of the master lexicon; (3) the need to strike the right balance between the master lexicon and rule sets so as to minimize the number of rules required, while making it generalizable to the different varieties being treated; (4) the lack of standardization that currently exists among Luso-African and Luso-Asian varieties, and the magnitude of variability that comes with describing non-standardized varieties; and (5) the lack of resources for measuring output quality (*see also* Section 4).

With LUPo , we can address (1) by augmenting the standard form of regular expression strings used in the original Unisyn Lexicon with advanced Perl constructs. The look-ahead and look-behind assertions enable LUPo to identify the appropriate vowel to insert between disallowed clusters in words such as *omnisciente* and *interregno*, and to anticipate the conditions for long-distance dissimilation, such as occurs among voiced stops in the words *obediência* and *dedozinho*.

Challenges (2) and (3) above should offer the benefit of hindsight to anyone interested in applying the Unisyn framework to other languages. For, while it may seem logical to start with the development of the accent-independent master lexicon before continuing on to develop the accent-specific rules, large-scale expansion of the master lexicon should only occur *after* sufficient efforts have been made to understand the basic characteristics of the varieties to be modeled. In the case of LUPo, which started with a small-scale master lexicon that was prematurely expanded, several iterations followed in which we aimed to properly address (3).

Challenges (4) and (5) above were known variables when setting about this work and remain LUPo's foremost hurdles. In terms of (4), which poses far fewer problems for better sourced varieties, we have applied informant frequencies in our decision making concerning the modeling of P_{Map}. Having collected detailed background information from our informants, we are also making use of the regional accent hierarchy (*see also* Section 1) to attribute certain allophonic alternations to a specific

speaker profile by introducing rules at the PERSON level that define the characteristics of a particular city zone. In general, while our methods for modeling regional varieties through direct contact with speakers introduces a level of variability that may be considered unwieldy, we regard variability as essential in creating rules for generating high-quality pronunciation output. By cataloging all observed variants with their context in the project's free, online, searchable databank, we achieve the added goal of establishing a separate resource for use by other speech technologies, such as ASR, as well as serving linguists.

Our approach to challenge (5) above is described in the section that follows.

4. Preliminary Results and Conclusions

In this paper, we presented the work of the LUPo project to create a rule-based pronunciation model for the resource-scarce variety of Portuguese spoken in the capital city of Mozambique, based on the Unisyn framework developed by Fitt [2].

Currently, LUPo consists of a 31,850-word master lexicon, and a conversion architecture with rules for generating accent-specific output for standard varieties from Lisbon, Rio de Janeiro, and São Paulo, as well as non-standard varieties representing the greater Maputo area, and the island of Catembe.

In terms of quality control, golden test sets have been developed for the three standard varieties based on hand-transcriptions by native speakers. From our modeling efforts, we also have control sets for the P_{Map} and Catembe varieties. For each of these – consisting of 177 hand-transcribed words, selected for adequate phonetic coverage – 100% accuracy was achieved in terms of generating output forms that matched the control set. However, beyond this preliminary quality control measure, the framework is limited in its capacity for verifying transcription quality over the entire set of pronunciations generated from the 31,850-word master lexicon. We aim to address this by: engaging language experts to validate output for the resource-scarce varieties, as part of the Common Orthographic Vocabulary project, or VOC, described in [7]; performing routine checks via LUPo's administrative back-end; and building a user feedback mechanism into LUPo's online Portal module.

In summary, although the Unisyn framework possesses no built-in methods for ensuring data integrity, supplementary measures can be taken to fine-tune the rules for varieties with few available resources. Overall, we hope to have shown that with careful attention to the staging of the master lexicon expansion and adequate handling of variability, the Unisyn Lexicon offers a potentially robust methodology for adaptation to other languages and resource-scarce varieties.

5. Acknowledgements

The authors gratefully acknowledge the support of the Fundação para a Ciência e a Tecnologia, and the cooperation of Dr. Susan Fitt, whose development of the original Unisyn Lexicon is the inspiration for this work. We are also extremely grateful to Inês Machungo and colleagues at the Eduardo Mondlane University.

6. References

1. Da Conceição, M.: A Brief Look at the Sociolinguistics of Ronga and Other Languages Spoken in Mozambique. Working Papers in Linguistics 16, 9-30 (1999)
2. Fitt, S.: Documentation and User Guide to UNISYN Lexicon and Post-lexical Rules. Online Technical Report, Centre for Speech Technology Research, http://www.cstr.ed.ac.uk (2000)
3. Wells, J.: The Accents of English. Cambridge University Press, New York (1982)
4. Fitt, S., Isard, S.: Synthesis of Regional English Using a Keyword Lexicon. In: Eurospeech Proc., pp. 823-826 (1999)
5. Rodrigues, M. C. M.: Lisboa e Braga: Fonologia e Variação. Fundação Calouste Gulbenkian, Lisbon (2003)
6. Gonçalves, P.: A Génese do Português de Moçambique. Imprensa Nacional – Casa de Moeda, Lisbon (2010)
7. Ferreira, J. P., Janssen, M., Barcellos de Oliveira, G., Correia, M., Müller de Oliveira, G.: The Common Orthographic Vocabulary of the Portuguese Language: A Set of Open Lexical Resources for a Pluricentric Language. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), pp. 1071-1075 (2012)

A quantitative study of disfluencies in formal, informal and media spontaneous speech in Spanish

Antonio Moreno Sandoval[1], Leonardo Campillos Llanos[1], Doroteo T. Toledano[2]

[1]Laboratorio de Lingüística Informática (LLI), Universidad Autónoma de Madrid,
28049 Madrid, SPAIN
{antonio.msandoval,leonardo.campillos}@uam.es

[2]ATVS Biometric Research Laboratory, Universidad Autónoma de Madrid,
28049 Madrid, SPAIN
doroteo.torre@uam.es

Abstract. A descriptive study of the prevalence of different types of disfluencies (fragmented words, restarts and vocalic supports) in spontaneous Spanish is presented based on a hand-annotated corpus. A quantitative account of differences among three types of registers (formal, informal and media) and several subtypes of text for each register is provided to analyze the importance of each disfluency class for a given register.

Keywords: Fragmented words, restarts, fillers, Spanish.

1 Introduction

The study of disfluencies is a key topic in current research. On the one hand, descriptive and psychological investigations are focused in the nature of the phenomena [2, 4-9, 13, 27, 28]. On the other hand, applied research in spontaneous speech processing is trying to detect and handle disfluencies [18]. For instance, the National Institute of Standards and Technology (NIST) has launched a series of competitive evaluations under the name Rich Transcription [19] in which spontaneous speech processing, and particularly disfluency detection, is one of the goals. In fact, disfluencies are one of the main problems in ASR [11].

Most of the research on disfluencies focuses in English language, although there are also studies for French, English, German, Italian, Japanese, Estonian or Mandarin Chinese [20-24] and also for European Portuguese [16, 29]; furthermore, there is some research on fillers performed with multilingual data [4]. Spanish has not been one of the languages with more research in this area, and few groups are conducting research in the field of spontaneous speech processing [25, 26]. Analogously, there are a few corpora available for spontaneous speech in Spanish. This study is based on the Spanish data from the C-ORAL-ROM corpus [17]. Previous research on disfluencies in this corpus has focused on acoustic-phonetic decoding of spontaneous speech [30] or has discussed the difficulties for the human transcribers [10]. The main con-

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

clusion of this last work [10] was that disfluencies are not the hardest problem for human transcribers, but the interaction features such as overlapping or speed rate. Another result was that media recordings are easier to transcribe than formal speech (talks in public context such as conferences, lectures, etc.). The hardest register to transcribe is, unsurprisingly, the informal, private speech. Another piece of research based on C-ORAL-ROM used also data from the MAVIR corpus [14], a collection of recordings and transcriptions gathered from professional conferences on language technologies and corporate presentations; in consequence, this research concentrated only in the formal register [3].

The goal of this paper is not a linguistic nor an acoustic description such as [3, 10, 30], but a quantitative analysis of three types of disfluencies (fragmented words, fillers and restarts) which may be of interest for ASR tasks. The choice of these types of phenomena is due to way they are encoded in the corpus, which makes them suitable for this brief analysis. Section 2 describes the C-ORAL-ROM corpus and each type of disfluency analyzed; and section 3 explains the results for every phenomenon.

2 Description of the C-ORAL-ROM Corpus

C-ORAL-ROM is a multilingual linguistic data bank that comprises four romance languages: Italian, French, Portuguese and Spanish. In this work only the Spanish sub-corpus [17], which contains around 300.000 spoken words, has been used. From a sociolinguistic point of view, speakers are characterized by their age, gender, place of birth, educational level and profession. From a textual point of view the corpus is divided into the parts shown on Table 1 [17]. This subdivision follows the design criterion of the C-ORAL-ROM corpus, which was mainly performed on the basis of register (formal and informal) [15]. A remarkable feature of the corpus design is the importance of the informal register, which means more than half of the corpus size.

In this study, in order to have a more balanced distribution between the main registers a large subset of the Spanish corpus has been selected, excluding telephone conversations and informal speech in public context (for example, narrations on a topic between unfamiliar speakers). Table 1 shows the subclasses considered for each of the three registers, along with the number of words in each of them. For the informal speech, the classification is based on the dialogic nature, and formal speech is divided between natural context (i.e. discourse recorded in a public setting: conferences, teaching, presentations, etc.) and media speech (or broadcast news speech). We have to point out that recordings from the media may not always be strictly classified in the formal register, since discourse from the sports programs and the talk shows may contain a quite informal speech.

An interesting feature of the C-ORAL-ROM corpus is that several types of disfluencies are annotated by hand. Trained linguists manually transcribed every recording, and along the orthographic transcription they marked disfluency phenomena following rigorous conventions [15]. Table 2 below summarizes and explains (with examples) the conventions used to mark every type of disfluency phenomena studied in this article.

Table 1. Distribution of words in C-ORAL-ROM (sum of words in all the files of each type).

Register		Type of texts	Words
FORMAL	Natural context	Preaching	12941
		Law	6203
		Political debate	6055
		Conference	12275
		Teaching	12353
		Prof. explanation	12063
		Business	9034
		Total	**70924**
	Media	Weather news	1591
		Interview	7640
		Reportage	35190
		Science	6108
		News	17335
		Sports	9330
		Talk-show	19976
		Total	**97170**
INFORMAL		Conversation	23495
		Dialogue	63590
		Monologue	41229
		Total	**128314**

Table 2. Transcription marks (with examples) for the disfluencies analyzed.

Mark	Disfluency	Meaning	Example
[/]	Simple retracting or restart.	Repetition or retrace.	no te hablo de [/] de [/] de equipos (emedsp02) ('I am not speaking to you of [/] of [/] of teams') esto facilita la acción [/] el éxito de la acción ('this facilitates the action [/] the success of the action') (enatbu03)
[///]	Retracting or restart.	Syntactic reformulation.	para &se [///] que te siga llamando (emedsp02) ('to &ke [///] for him to keep on calling to you')
&	Before a fragmented or truncated word.	A non-complete word (self-correction).	trabajamos mucho para &Sudamer [/] Sudamérica ('we work a lot for South &Am [/] South America') (enatpe01)
&eh &ah &mm	Vocalic support or filler.	The speaker uses it to keep his / her turn.	saben / &eh / cómo / puede mejorar su producto ('they know / er / how / their product can improve') (enatbu03) es algo / &mm / muy interesante (enatps01) ('it is something / um / very interesting')

Although an inter-annotator agreement test was not performed, every transcription was reviewed by another linguist to guarantee the accuracy of the data.

3 Results

In order to estimate the prevalence of the three types of disfluency in each register and subclass, the procedure used has been:

1. Count the number of marks for each kind of disfluency in every text.
2. Calculate the ratio between number of words in each text and the number of occurrences of a given mark in the text (i.e. the average number of words by disfluency in each recording).
3. Calculate the average ratio for the texts in the same subclass and register.

Results are shown numerically on Table 3 and graphically in Figures 1 to 3. The prevalence measure used is inversely related to the frequency of the phenomenon in that subclass: a higher number means that the disfluency is less frequent and vice versa.

Table 3. Average number of words between occurrences of each type of disfluency analyzed.

Register		Type of texts	Truncated	Supports	Rep./Restart
FORMAL	Natural context	Preaching	468,69	146,54	471,33
		Law	240,49	130,04	97,53
		Political debate	206,06	70,36	123,13
		Conference	138,93	134,29	91,82
		Teaching	129,42	63,74	45,90
		Prof explanation	97,27	68,99	35,78
		Business	118,27	40,54	45,33
		Average	**199,88**	**93,50**	**130,12**
	Media	Weather news	0,00	554,00	259,50
		Interview	155,56	31,00	44,85
		Reportage	605,34	79,02	118,77
		Science	202,94	55,23	78,45
		News	619,35	233,42	224,88
		Sports	124,06	102,86	58,80
		Talk-show	187,80	172,14	59,61
		Average	**270,72**	**175,38**	**120,69**
INFORMAL		Conversation	339,87	492,55	50,89
		Dialogue	133,17	406,14	42,64
		Monologue	190,10	120,85	40,01
		Average	**221,05**	**339,84**	**44,51**

Figure 1: Distribution of the **average number of words per fragment** in the different types of texts.

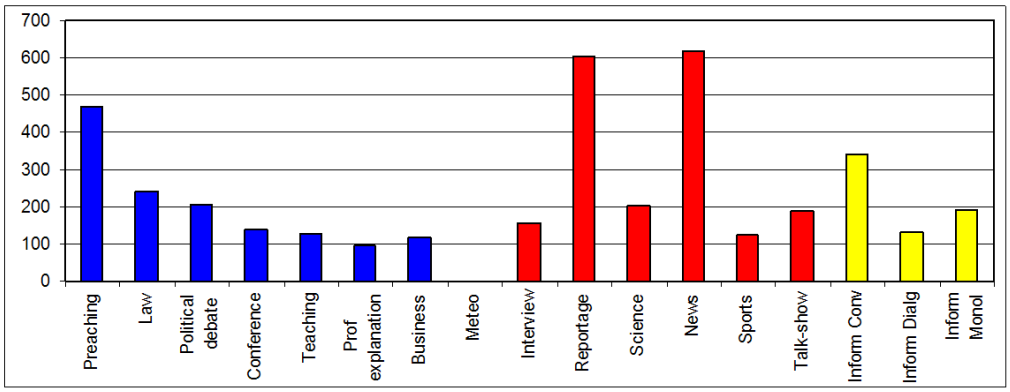


Figure 2: Distribution of the **average number of words per restart** in the different types of texts.

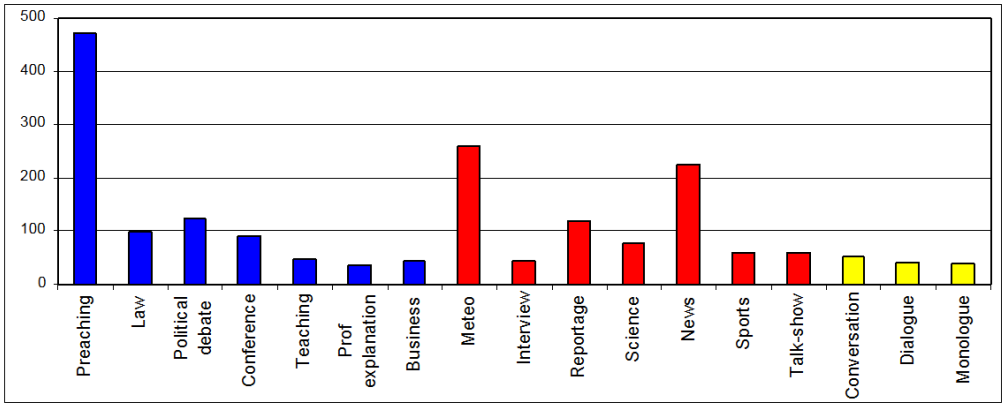
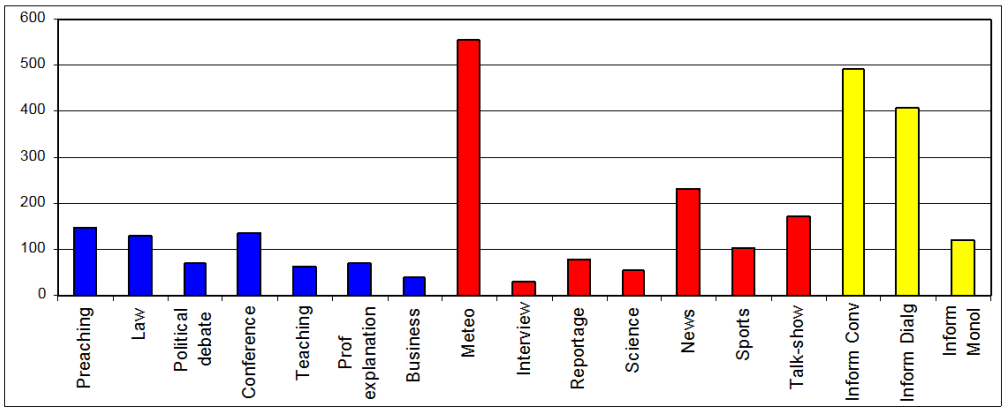


Figure 3: Distribution of the **average number of words per support** in the different types of texts.



3.1 Fragmented words

A fragmented (or truncated) word is an incomplete pronunciation of a given lexical item. Truncated words abound in spontaneous speech since speakers are thinking while they are speaking, and they retrace and self-correct when a segment is already uttered. According to a still unpublished research performed in a phonological lexicon of over 400,000 items, in Spanish the most frequent number of syllables for a full inflected word is 4, followed by 5 and 3. The number of monosyllables and bi-syllables in the same Spanish lexicon is less than 20,000 (only 5% of the words). That could explain that this disfluency is relatively common in Spanish.

Results for fragmented words are shown in both Table 3 and Figure 1. Fragmented words are a very rare in media texts, especially in news and reportages, probably because of their almost-read nature. In the meteorological reports there are no occurrences. The homilies (non-read commentaries on the Holy Writings during the mass) are also a subclass in the formal register, where the truncated words are very infrequent. In contrast, they abound in several formal subclasses (professional explanations, business, teaching and conferences) as well as in informal dialogues.

3.2 Restarts and repetitions

The retracting is the most frequent fragmentation phenomenon in spontaneous speech; in the scientific literature, they have also been called *retrace and repair sequences* [1], *restarts* or *false starts* [15], or *repairs* [18]. Actually, the term *reparandum* is frequently used to refer to the sequence of syllables or words (or even the beginning of an utterance) that is deleted, while the replacing fragment is called *repair* [12, 27]:

un / interesante debate / que también **vas** [/] **vamos** a tratar (enatte01)
REPARANDUM REPAIR
(‘an / interesting debate / which you are also going to [/] we are going to discuss’)

The speaker hesitates while trying to find the best way to express himself and retracts his speech before choosing between two alternatives. The retracting phenomenon is almost always accompanied by the repetition (complete or partial) of the linguistic material and clearly causes a loss of the informational value of the retracted material, which is abandoned by the speaker in favor of the chosen alternative [15].

In our analysis, we counted together both the restarts and the repetitions (or repeats, as they are also referred to in some articles [7]). The most frequent repetitions and retraces are short items (usually just one word) and mainly mono- or disyllabic (as a matter of fact, prepositions and determiners the majority of the times) [3].

As in the previous case, restarts and repeats are very rare in preaching, weather forecast and broadcasting news. The retracting appears very often in informal speech, but also in several subclasses of formal and media registers: professional explanation, teaching and business, on formal speech; and interviews, sports and talk shows, on the media domain. In all those subclasses, its interactive nature of the process could explain why speakers hesitate so frequently.

3.3 Vocalic supports

Vocalic supports (or filled pauses) are a paralinguistic element used for helping the speaker to think about the way the next utterance must be expressed. In the literature, they are also called *fillers* [6, 7], *pause fillers* or *hesitators* [1]. In ASR of spontaneous discourse, fillers tend to be included in the grammar module along with full lexical words [12]; thus, it is important is to model filled pauses in a way that they would not be erroneously recognized. For example, Spanish *eh* may be mistaken with the conjunction *e* (used as a variant of *y* before a syllable beginning with *i-*/*hi-*) or *ah* may be interpreted as the preposition *a* (and vice versa). According to previous research in Spanish formal speech [3], the most frequent filler is *eh,* whereas *ah* is scarce.

As shown in Figure 3, the filled pauses are typical of the formal speech, where the speaker wants to be precise in his or her speech. On the other side, vocalic supports are scarce in informal speech, where the speed of the interaction and a more relaxed situation do not favor the filled pauses. In media recordings, those subclasses closer to the informal speech (talk shows and sport) have less average number of vocalic supports. Weather and broadcast news subclasses are the ones with less number of filled pauses, mainly because their near-read nature. Some research based on Spanish data from formal register [3] has stated that the frequency of filled pauses in every subregister may be influenced by other factors different from the dialogic style or the genre: for instance, the speaker's speaking style, the difficulty related to the degree of complexity of the contents or the anxiety induced by the communicative situation.

4 Conclusions

The present empirical study on the prevalence of tagged disfluencies in a spontaneous speech corpus of Spanish shows interesting quantitative data about the frequency and distribution among subclasses and registers of three important types of disfluencies. Retractings and vocalic supports are more frequent than fragmented words in all the types of recordings. These results, based on data from informal and formal register, match up with the results obtained from formal data in previous research [3]. Attending only disfluency type and prevalence, instead of register, it would be more accurate to divide the corpus into the following classes that share similar disfluency rates:

- Professional explanation, teaching, business, interviews, sports and talk shows present a high frequency of retractings.
- Talk shows and sport news present a similar, high rate of vocalic supports to that in informal speech.
- Preaching speech, weather and broadcast news do not abound with truncated words

The results of this study could be interesting in the design of ASR tasks, since they provide a measure of the disfluency rates per type of disfluency found in Spanish for different types of speech. This will allow researchers in the field of ASR to focus on the most important types of disfluency for the particular type of speech they have to process.

References

1. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Longman Grammar of Spoken and Written English. London: Longman. (1999)
2. Boula de Mareüil, P., Habert, B., Bénard, F., Adda-Decker, M., Baras, C., Adda, G., Paroubek, P.: A quantitative study of disfluencies in French broadcast interviews. In Proc. ISCA Workshop on Disfluency in Spontaneous Speech, DiSS 2005 (2005)
3. Campillos, L., Alcántara, M.: Speech Dysfluencies in Formal Context. Analysis based on Spontaneous Speech Corpora. In Proc. Corpus Linguistics Conference 2009. (2009)
4. Candea, M., Vasilescu, I., Adda-Decker, M.: Inter- and intra-language acoustic análisis of autonomous fillers. Proceedings of DiSS'05, Disfluencies in Spontaneous Speech Workshop 2005, Aix-en-Provence (2005)
5. Campione, E., Véronis, J. Pauses and hesitations in French spontaneous speech. Proc. ISCA Workshop on Disfluency in Spontaneous Speech, DiSS 2005. (2005)
6. Clark, H. Speaking in time. Speech Communication, 36, pp. 5-13 (2002)
7. Clark, H. & Wasow, T. Repeating Words in Spontaneous Speech. Cognitive Psychology, 37(3), pp. 201-242 (1998)
8. Ferreira, F., Lau, E. F., Bailey, K. G. D.: Disfluencies, language comprehension, and Tree Adjoining Grammars. Cognitive Sciences, 28, pp. 721-749 (2004)
9. Gilquin, G., De Cock, S.: Errors and Disfluencies in Spoken Corpora. Special issue of International Journal of Corpus Linguistics, 16:2. Amsterdam: John Benjamins (2011)
10. González Ledesma, A., De la Madrid, G., Alcántara Plá, M., De la Torre, R., Moreno-Sandoval, A.: Orality and Difficulties in the Transcription of Spoken Corpora. In Proc. of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, 2004, Lisbon (2004)
11. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Thory, Algorithm and System Development. NJ: Prentice Hall PTR. (2001)
12. Jurafsky, D., Martin, J.: Speech and Language Processing. 2nd edition. Prentice Hall Series in Artificial Intelligence. (2008)
13. Levelt, W.: Speaking: from intention to articulation. MIT Press. (1989)
14. MAVIR Corpus. Distributed by the Fundación General de la Universidad Autónoma de Madrid (FGUAM) (2008)
15. Moneglia, M.: The C-ORAL-ROM resource. In Cresti, E., Moneglia, M. (eds.) p. 27 (2005)
16. Móniz, H., A. I. Mata, C. Viana: On filled pauses and prolongations in European Portuguese. Proc. INTERSPEECH – ISCA, August 27 - 31, pp. 2645-2648. (2007)
17. Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, JM., de la Torre, R.: The Spanish Corpus. In Cresti, E., Moneglia, M. (eds.) C-ORAL-ROM: Integrated reference Corpora for Spoken Romance Languages, pp. 135-161. Amsterdam: John Benjamins (2005)
18. Nakatani, C. H., J. Hirschberg: A Corpus-based study of repair cues in spontaneous speech. Journal of the Acoustic Society of America, 95:3, pp. 1603-1616. (1994)
19. NIST Rich Transcription Evaluation Homepage, http://www.itl.nist.gov/iad/mig/tests/rt/.
20. Proceedings of the ICPhS Satellite meeting on Disfluency in Spontaneous Speech, UC Berkeley DiSS-1999. (1999)
21. Proc. DiSS-2001, Disfluencies in Spontaneous Speech Workshop 2001, August 29-31, Edinburgh, Scotland, UK. http://www.isca-speech.org/archive_open/diss_01/ (2001)
22. Proc. DiSS-2003, Disfluencies in Spontaneous Speech Workshop 2003, September 5-8, Göteborg, Sweden. http://www.isca-speech.org/archive_open/diss_03/ (2003)

23. Proc. DiSS-2005, Disfluencies in Spontaneous Speech Workshop 2005, Aix-en-Provence. http://www.isca-speech.org/archive_open/diss_05/ (2005)
24. Proc. DiSS-2010, 5th Workshop on Disfluency in Spontaneous Speech, Tokyo. (2010)
25. Rodríguez, L. J., Torres, I., Varona, A.: Annotation and Analysis of Disfluencies in a Spontaneous Speech Corpus in Spanish. Proc. ISCA Workshop on Disfluency in Spontaneous Speech, DiSS 2001. (2001)
26. Rodríguez, Luis J., Torres, I.: Spontaneous Speech Events in Two Speech Databases of Human-Computer and Human-Human Dialogs in Spanish. Language and Speech, 49, 3, pp. 333-366 (2006)
27. Shriberg, E.: Preliminaries to a Theory of Speech Disfluencies. U. Cal. Berkeley. Ph.D. Thesis. (1994)
28. Shriberg, E.: Spontaneous Speech: How People Really Talk and Why Engineers Should Care. Proc. EUROSPEECH 2005 - INTERSPEECH 2005, Lisbon, Portugal, pp. 1781-1784. (2005)
29. Silva, A.: Caracterização segmental e prosódica de disfluências em discurso espontâneo. Master's Thesis. Universidade de Aveiro (2006)
30. Torre Toledano, D., Moreno Sandoval, A., Colás Pasamontes, J., Garrido Salas, J.: Acoustic-phonetic decoding of different types of spontaneous speech in Spanish, In Proc. DiSS-2005, Aix-en-Provence, pp. 165-168. (2005)

A Computational Morphology for Spanish Sign Language Nominal Inflection

Jordi Porta, Fernando López-Colino, and José Colás

Escuela Politécnica Superior.
Universidad Autónoma de Madrid.
Av. Francisco Tomás y Valiente, 11.
28049 Madrid, Spain.
{jordi.porta,fj.lopez,jose.colas}@uam.es

Abstract. This paper presents a finite-state computational model for Spanish Sign Language nominal morphology with particular attention to the treatment of morphological alternations. A computational morphology consists of a lexicon, a rewrite rules component relating lexical representations to surface forms, and a morphotactic component. All these components are compiled into finite-state transducers. Nominal morphology in Spanish Sign Language deals with bound morphemes through inflection, introflection and suppletion. The model and its implementation deal with plurals, numeral incorporation and irregularity and can be extended to other word classes as well to other sign languages. The implemented technology allows the application of the morphology either to the analysis or to the generation of signs.

1 Introduction

Two central aspects for a morphology are word formation and morphological alternations. Word formation is governed by the principles of combinations of stems and affixes. Morphological alternations are the different realizations of a morpheme depending on its phonological environment and the composition of the word. Sign languages (SLs) have a wide range of morphological processes, but this paper deals mainly with nominal inflection in Spanish Sign Language (LSE), leaving aside derivation, compounding and classification as well other word classes. The kind of bound morphemes found in LSE are flective and introflective. Flective morphemes are concatenated to the phonological form, like a movement or, in our setting, a repetition of the original movement, sometimes in a direction. Introflective morphemes are non-concatenative and cause internal changes to the phonological form of the sign. LSE has alse irregularities that can be modeled using suppletive base forms, i.e., different and unrelated base forms for the same sign, bearing different grammatical values. Finally, modulation, i.e., the modification of suprasegmental elements like the stress or the tone in spoken languages, is also present in LSE as well as in other SLs.

2 Transcription of Sign Languages

While writing systems are intended for the purpose of communication, transcription systems are used for specialists to record or analyze the form of utterances. A faithful reproduction of an utterance is only obtained by audio recording, in the case of speech acts, or by video recording, in the case of signed productions. English writing and SignWriting are examples of writing systems, but IPA (International Phonetic Alphabet) and HamNoSys (Hamburg Notation System) are transcription systems. HamNoSys is a lineal transcription system for SLs developed at the University of Hamburg. HamNoSys was designed as a phonological or phonetic transcription system using highly iconic symbols.[1] HamNoSys has evolved to its current fourth version [18]. A sign's transcription consists of symmetry operator, initial configuration and actions. In turn, an initial configuration consists of handshape, hand orientation and location. The encoding scheme is the following: Symmetry (S) + Handshape (H) + Orientation (O) + Location (L) + Movement (M), but only H and O, which are called configuration (G), are mandatory. For the encoding of non-manual components (NMCs), there are used other tiers and an inventory of descriptions. Mouth pictures are described as visemes encoded using the SAMPA alphabet.

It is worth noting that HamNoSys is ambiguous in the sense that a sign can be expressed in a variety of equivalent forms from the point of view of performance. This ambiguity poses difficulties to automatic processing since this formal variation should be taken into account. However, without loss of generalization, we will suppose in this paper that signs are transcribed according to some criteria easing their processing. These criteria will be presented when appropriate in further sections.

3 Previous Formalization of Sign Language Morphology

The phonological models of how signs are composed begins historically with the Move-Hold model [13], which borrows some ideas from auto-segmental phonology, and continues with the Hand Tier model [17], which attempts to improve the previous model. However, the use of HamNoSys is advocated in [11] for the interlinear transcription and glossing of SL data, representing an inflected sign by its base form plus its morphological variations separated with bars.

Much of the previous work done in SL processing, including computational morphology, has been carried on within the ViSiCAST Project [19]. How the sign's morphological information can be abstracted using λ-calculus is shown in [6]. In the lexicon, reference forms could be represented as lambda expressions like $\lambda x.$ ¨ ○ ≏ ≦ ˇ x ˘ ↓, where location has been abstracted. However, the number of obligatory and optional morphemes that a sign can admit makes λ-expressions

[1] LSE has its own transcription system called SEA (*Sistema de escritura alfabética*) [7]. The system was designed for being easily employed with computers and uses the Latin alphabet. A successful conversion of SEA transcriptions to HamNoSys had been reported [16].

difficult to interpret. Consequently, the previous model was replaced by another model in which phonological information is split up into parameters and represented with feature structures within the framework of the Head-driven Phrase Structure Grammar (HPSG). In this framework, linguistic information at different levels, including the phonological level, is co-indexed. Typed unification is used as the operation for composing morphological and lexical information, which can be finally mapped into a HamNoSys sequence. Continuing with the work started in ViSiCAST, an HPSG sign language generation component of an English-to-BSL prototype system is presented in [14] and [4].

A finite-state morphological analyzer for modeling some of the verbal aspectual distinctions in American Sign Language (ASL) is presented in [20]. In that approach, signs are not represented by transcriptions in any formalism but by a more abstract formal representation not committed to a particular underlying phonological form of the sign. In this representation, each parameter is represented with a transparent and comprehensible shorthand system. The ASL verb SEE is represented as <Type:1H DH:Vin NDH:none Loc:face −Touch −Twist −Redup +Arc −Slow>, which represents a one-handed (Type:1H) sign with a dominant hand (DH) with a handshape with the index and middle fingers extended (V) and the palm oriented towards the signer (in). The movement begins at the height of the face (Loc:face) without touching it, and the movement continues with a normal speed (−Slow) describing and arc (+Arc). The path movement is reduplicated (+Redup). In ASL, as in other SLs, aspectual distinctions affect mainly to movement and NMCs. Only the two most common aspectual values, habitual and continuative, are considered in the paper. A rewrite rule for the habitual aspect is expressed as $(- \rightarrow + \ / \ .^*$ Aspect:Hab $.^*$ ___ Redup|Arc). However, some unpronounceable verbal forms are generated when some aspect values are added. These co-articulation constraints are dealt with a series of transducers composed together in a cascade correcting forms appropriately.

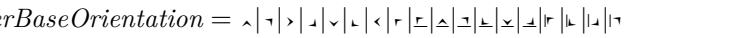
4 Formalization

In a computational morphology, one want to relate the lexical level, consisting of base forms with abstract linguistic descriptions in the form of feature sets, to surface realizations. Traditional phonological grammars were formalized in [3] using ordered sequences of rewrite rules that converted abstract phonological representations into surface forms through a series of intermediate representations. Such rewrite rules have the general form $(\alpha \rightarrow \beta \ / \ \gamma \ ___ \ \delta)$ where α, β, γ, and δ can be arbitrarily complex strings or feature-matrices. In [10], it was established that such rewrite rules describe regular relations when α, β, γ and δ are regular expressions. Regular relations can be implemented as finite-state transducers. As a consequence, an ordered series or cascade of rules can be compiled to FSTs and composed to obtain a single FST representing the entire rule series. These results were rediscovered and put into a more algebraic form by [12], who developed the techniques for the compilation of rewrite rules to FSTs. The notation used here for regular expression operators is based on POSIX Extended

Regular Expressions. The symbol 'ϵ' denotes the empty string. The operator '$\circ$' is used for the composition of two rational expressions. Direct or conditional rewrite rules are applied left-to-right with a longest match strategy.

Lexical base forms are conventional forms that usually coincide with dictionary headwords. However, these conventions are not universal. For nouns in SL dictionaries, the base form is expected to correspond to the singular form of a sign. Lexicon entries are classified according to their inflectional paradigm, i.e., the way the base form should be manipulated in order to generate all the desired morphologically-related sign forms. At lexical level, a sign's base form is phonologically represented by its HamNoSys transcription and its morphological features, which are represented by conventional symbols like '-SG' indicating 'singular' number value. These features are given in a fixed order. At surface level, the sign's wordform is only represented by its HamNoSys transcription.

In order to formalize the morphology, let be the alphabet Σ the union of set of HamNoSys symbols and the set of morphological tags, and let be S, H, O, L and M regular languages over Σ describing each of the string values of each of the phonological parameters of signs. Let be *Tags* the set of non-empty sequences of morphological tags. To put an example, for O, its language is defined by the following set of regular expression definitions:

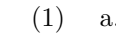
$FingerBaseOrientation =$ 
$FingerOrientation = FingerBaseOrientation\ FingerBaseOrientation?$
$PalmOrientation =$
$O = FingerOrientation\ PalmOrientation$

5 Nominal Inflective Plurals

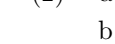
In LSE, numerals or indefinites are added after singular nouns to express number, as in the examples in BROTHER THREE ('three brothers') and CAR SOME ('some cars'). However, repetition is the most iconic representation of number and, in almost all SLs, nominal signs and classifiers can add some kind of repetition to their singular form to express frequency, abundance or descriptive plurals. The use of repetition in nouns for number inflection depends on semantic and phonological constraints [5, 9]. In order to admit repetition, a sign should be one-handed and signed in the neutral space, two-handed symmetrical or two-handed with a passive hand. Body-anchored signs and many signs including repetition do not use to inflect for number. Normally, adding a repetition to a sign means to perform the sign twice, but when repetition is included in the sign's movement, the sign should be performed three times. As it will be seen in this section, number inflection has many variations. In order to implement them, the following morphematic tags, which account for the different realizations of number, are proposed: PL.REP (in-place repetition), PL.REP.C (in-place repetition with circular joining movement), PL.REP.H (horizontal directional repetition), PL.REP.V (vertical directional repetition), PL.SLID.H (horizontal directional sliding), PL.SLID.V (vertical directional sliding), PL.BIMAN.H

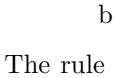
(two-handed rep. with alternating horiz. movement), and PL.BIMAN.V (two-handed rep. with alternating vert. movement).

Signs like NIGHT or SYMBOL express abundance using **in-place repetition**, as it can be seen respectively in examples (1) and (2), but repetition is applied differently. For NIGHT, which is a two-handed symmetrical sign, repetition affects the performance of the two hands, but for SYMBOL, which is a two-handed asymmetrical sign with a passive hand, repetition applies only to the active hand.

(1) a.  (NIGHT-SG)

b. (NIGHT-PL.REP)

(2) a.  (SYMBOL-SG)

b.  (SYMBOL-PL.REP)

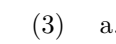
The rule *RepeatInPlace* (1) implements the in-place repetition of signs dealing with one-handed and two-handed signs. It first applies the rule *MarkMovement* (2) that groups movement symbols and marks the point of insertion of the repetition operator. This marking depends on how hands are used. Only the movements of the active hands are marked. Once the movement is marked, *InsertInPlaceRepetition* (3) replaces the marking symbol by a repetition symbol.

$$RepeatInPlace = MarkMovement \circ InsertInPlaceRepetition \tag{1}$$

$$\begin{aligned} &MarkMovement = \\ &(\epsilon \rightarrow {}^{(} \ / \ L \ ___ \ M) \circ (\epsilon \rightarrow {}^{)}{>} \ / \ L^{(}M \ ___ \ \text{Tags}) \circ \\ &(\epsilon \rightarrow {}^{(} \ / \ {}^{[} \ ___ \ M_{\swarrow}) \circ (\epsilon \rightarrow {}^{)}{>} \ / \ L^{[(}M \ ___ \ {}_{\swarrow}) \end{aligned} \tag{2}$$

$$InsertInPlaceRepetition = (> \rightarrow {}^{+}) \tag{3}$$

Other signs express its frequency plural joining each of the in-place repetitions by means of a circular movement, like MONDAY, as it can be seen in the example (3b).

(3) a.  (MONDAY-SG)

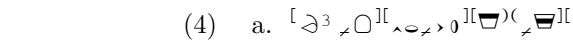
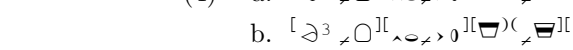
b. (M.-PL.REP.C)

Rules dealing with this kind of repetition are:

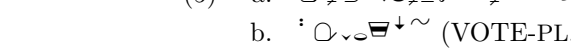
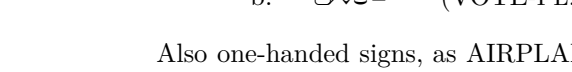
$$RepeatCircular = MarkMovement \circ InsertInPlaceCircularRepetition \tag{4}$$

$$InsertInPlaceCircularRepetition = (> \rightarrow {}^{\circlearrowleft}) \circ MarkMovement \circ (> \rightarrow {}^{+}) \tag{5}$$

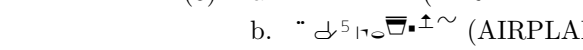
Many nouns that are repeated to express frequency are deverbal nouns. Consequently, their plural forms do not differ from the corresponding verbal signs displaying a frequentative aspect value. Two examples of deverbal nouns are LETTER and VOTE. The first case, LETTER is derived from the verbal sign with the meaning of 'sending letters' and is repeated in-place, as is shown in example (4).

(4) a.  (LETTER-SG)
b.  (L.-PL.REP.C)

However, as in can be seen in example (5), VOTE-SG is articulated as a two-handed asymmetrical sign with the passive hand representing iconically an urn, but its frequentative plural is realized using both hands as articulators, alternating their movement when the sign is repeated.

(5) a.  (VOTE-SG)
b.  (VOTE-PL.BIMAN)

Also one-handed signs, as AIRPLANE, turns two-handed to express its plural, but unlike VOTE, which is symmetrized in the vertical plane, the symmetry for AIRPLANE is horizontal, see example (6b).

(6) a. (AIRPLANE-SG)
b.  (AIRPLANE-PL.BIMAN.H)

The following rewrite rules are applied to singular base forms and making signs two-handed symmetric alternating hands horizontally in repetition. Because the final location of a movement can be determined in HamNoSys either by the direction and size of the movement path or by specifying it as a target location, the parameter M is decomposed as $M = RT$, where R represents all the relative movements and T the target locations. Note that absolute movements specifying hands as target locations are removed with rule (7).

$$\begin{aligned} &\mathit{MakeAlternatingHandsHorizontally} = \mathit{RemoveTargetLocation} \circ \\ &\quad \mathit{ProjectDominantHand} \circ \mathit{InsertParallelSymmetry} \circ \\ &\quad \mathit{MarkMovement} \circ \mathit{InsertAlternatingMovement} \end{aligned} \tag{6}$$

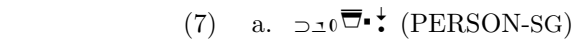
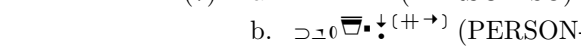
$$\mathit{RemoveTargetLocation} = (T \rightarrow \epsilon \ / \ [R \ ___ \ \measuredangle M]) \tag{7}$$

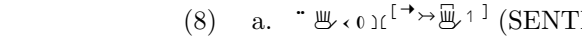
$$\mathit{InsertParallelSymmetry} = (\epsilon \rightarrow \ ^{:} \ / \ \hat{} \ ___ \ ) \tag{8}$$

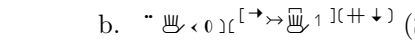
$$\mathit{InsertAlternatingMovement} = (> \rightarrow \ ^{\sim}) \tag{9}$$

$$\begin{aligned} &\mathit{ProjectDominantHand} = (S \rightarrow \epsilon) \circ (H|O|L|M \rightarrow \epsilon \ / \ \measuredangle \ ___ \ ^{]}) \circ \\ &\quad (^{[} \rightarrow \epsilon \ / \ ___ \ (H|O|L|M) \measuredangle ^{]}) \circ (\measuredangle ^{]} \rightarrow \epsilon) \end{aligned} \tag{10}$$

Directional repetition is used to express abundance plurals and is applied to objects represented mentally as spatial-ordered series Default direction of repetition is horizontal, going from left to right, as it can be seen for PERSON in example (7), but in other cases, direction is vertical, from the height of the baseform to bottom, as it can be seen in example (8) for SENTENCE.

(7) a.  (PERSON-SG)
b.  (PERSON-PL.REP.DIR.H)

(8) a. 

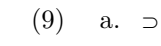
b.  (SENTENCE-PL.REP.DIR.V)

Horizontal or vertical directional repetition is orthogonal to the sign's own direction. Horizontal repetition is realized with the following rules and vertical repetion is realized in a similar way.

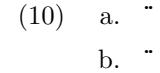
$$RepeatHorizontally = MarkMovement \circ InsertHorizontalRepetition \quad (11)$$

$$InsertHorizontalRepetition = (> \rightarrow {}^{(\#\rightarrow)}) \quad (12)$$

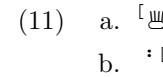
Sometimes, directional repetition can become into a **sliding movement**, like in the plurals of TREE, HOUSE or WORD, which can be seen in the following examples:

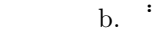
(9) a.  (WORD-SG)

b. (WORD-PL.SLID)

(10) a.  (HOUSE-SG)

b. (HOUSE-PL.SLID)

(11) a.  (TREE-SG)

b.  (TREE-PL.SLID)

Horizontal sliding is implemented with the following rules:

$$SlideHorizontally = InsertHorizontalSlide \quad (13)$$

$$InsertHorizontalSlide = (\epsilon \rightarrow {}^{\rightarrow} \;/\; {}^{[}M \;___\; {}_{\swarrow}) \circ (\epsilon \rightarrow {}^{\rightarrow} \;/\; {}_{\swarrow}M \;___\; {}^{]}) \circ (\epsilon \rightarrow {}^{\rightarrow} \;/\; M \;___\; Tags) \quad (14)$$

Additionally, the plural of abundance can adopt a form of a **distributive plural**. In this kind of plural, the sign is repeated several times and each instance is placed in a particular location, as in 'TREE-LOC.LEFT TREE-LOC.CENT TREE-LOC.RIGHT'. The location of the sign, which should be locatable, is changed with rewrite rules as the following:

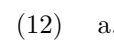
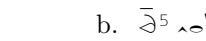
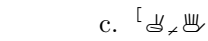
$$MakeLocationRight = (\epsilon \rightarrow {}^{\blacksquare} \;/\; L \;___\; M?\; Tags) \quad (15)$$

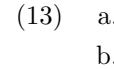
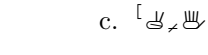
Finally, when repetition is used to emulate the arrangement of objects in the real world, using the sings or their classifiers, it is said that the plural is a **descriptive plural**. This kind of plural is not considered properly a morphological form of plural. It is worth noting that there exist a continuum between systematic or morphological plurals and no systematic or descriptive plurals, as evidenced by directional plurals, which are systematic and descriptive at the same time.

6 Nominal Introflective and Suppletive Plurals

A few temporal nouns in LSE admit numeral incorporation by introflecting the configuration of the number into the sign. These nouns correspond to the names of time units: MINUTE, HOUR, DAY, WEEK, MONTH and YEAR. They all admit numerals up to ten, but apart from the configuration, when numerals are incorporated, some other phonological parameters of the sign are modified, so that, in some cases, introflection takes place in a suppletive form of the sign. Suppletive forms of plural exist in some LSE communities or in some registers for the temporal nouns HOUR, MONTH and WEEK. Suppletion supposes the substitution of the neutral base form of the sign by another base form conveying some grammatical values, which in the case of temporal nouns, the plural value.

In order to show how suppletion and inflection interact, lets suppose a language variety in which DAY and YEAR have suppletive forms of plural. In the case of DAY, the singular form, glossed by the sign DAY-1, and the incorporation of three, glossed by DAY-3, are shown in example (12). For YEAR, examples are given in (13).

(12) a.  (DAY-1)
b.  (DAY-3)
c.  (DAY-7)

(13) a.  (YEAR-1)
b. (YEAR-2)
c.  (YEAR-7)

The replacement of the phonological base form in the case of signs with suppletive plural is realized by the rules in (16), which implements a list of signs with its corresponding plural suppletive form.

SuppletivePlurals =
($YEAR \rightarrow YEARS$ / ^ ___ (-2|-3| . . . |-10)) ∘
($DAY \rightarrow DAYS$ / ^ ___ (-2|-3| . . . |-10)) (16)

For the incorporation of numerals above five, its realization in assumed to be parallel to the construction of numerals. The configuration of FIVE is incorporated into the passive hand and the active hand incorporates the numeral resulting of the subtraction of five from the original numeral. Rules in (17), (18) and (19) are responsible for the incorporation of numerals in one-handed signs.

IncorporateNumeral =
IncorporateNumeralToDominantHand ∘
IncorporateNumeralToNonDominantHand (17)

$$\begin{aligned}&\textit{IncorporateNumeralToDominantHand} = \\ &\quad (H \rightarrow \text{☝} \;/\; \hat{}S? \;___\; O.^*(\text{-}2|\text{-}7)) \circ (H \rightarrow \text{✌}^5 \;/\; \hat{}S? \;___\; O.^*(\text{-}3|\text{-}8)) \circ \\ &\quad (H \rightarrow \text{🖖} \;/\; \hat{}S? \;___\; O.^*(\text{-}4|\text{-}9)) \circ (H \rightarrow \text{🖐} \;/\; \hat{}S? \;___\; O.^*(\text{-}5|\text{-}10))\end{aligned} \tag{18}$$

$$\begin{aligned}&\textit{IncorporateNumeralToNonDominantHand} = \\ &\quad (\epsilon \rightarrow \lceil \;/\; \hat{}S? \;___\; H.^*(\text{-}6|\ldots|\text{-}10)) \circ (\epsilon \rightarrow \text{↙🖐}\rfloor \;/\; \lceil H \;___\; O.^*(\text{-}6|\ldots|\text{-}10)) \circ \\ &\quad (\epsilon \rightarrow \lceil \;/\; O \;___\; L\;.^*\;(\text{-}6|\ldots|\text{-}10)) \circ (\epsilon \rightarrow \text{↙▪🤚}\rfloor \;/\; \lceil L \;___\; M? \;(\text{-}6|\ldots|\text{-}10))\end{aligned} \tag{19}$$

7 Implementation

A compiler for regular expressions and rewrite rules to finite-state automata and transducers has been implemented using the OpenFST library [1]. The compilation of rewrite rules uses the algorithm given in [15]. Symbols are taken from SiGML (Signing Gesture Markup Language), a XML application based on HamNoSys that provides, among other things, HamNoSys symbol names for the encoding of manual features. For the non-manual component, SiGML includes different tiers for shoulders, body, head, eye-gaze, facial expression and two subtiers for mouthing. SAMPA codes are used for speech mouthing. SiGML is the input notation used by the JASigning application, a synthetic sign language performance system superseding the earlier software developed in ViSiCAST and eSIGN projects. Rules as those seen in previous sections are written as the following:

```
{RemoveTargetLocation} =
  ({T} -> <eps> / <hamparbegin> {R} _ <hamplus> {M} <hamparend>) .o.
  ({T} -> <eps> / <hamparend>? {R} _ <hamparbegin>? {Tags}* $) ;
```

8 Conclusion

The implementation of the morphological model presented here is difficult to evaluate providing that LSE, and SLs in general, are under-resourcered languages. To our knowledge, there are no HamNoSys texts in LSE that could be analyzed with the lexicon to measure its coverage, but since one of the aims of these work is to build morphological generation technologies for SLs to be used in machine translation systems, a lexicon based on examples in [8] had been tested using avatar technology in order to guide the implementation of the theory. However, even a human sign recognizing evaluation is far from being useful. From one hand, some errors come from the performance of the avatar. From the other hand, other recognizing error can come from the modeling of phenomena in languages in process of standardization.

According to [2], all known phonological and morphological processes, organized by different theories in different ways, can be treated as regular relations definable in terms of regular expressions. This article has shown how this result is applied to the kind of morphological phenomena around LSE nouns. The

proposed model could be elengantly extended to cover other word classes as well as to other SLs. It could be applied to morphological sign generation in machine translation or to automatic sign recognizing or analysis of HamNoSys transcriptions.

References

1. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., Mohri, M.: OpenFst: A general and efficient weighted finite-state transducer library. In: Proceedings of the 12th Int. Conference on Implementation and Application of Automata, (CIAA 2007), LNCS. pp. 11–23. Springer, Prague, Czech Republic (2007)
2. Beesley, K., Karttunen, L.: Finite State Morphology. CSLI, Palo Alto (2003)
3. Chomsky, N., Halle, M.: The Sound Pattern of English. Harper and Row (1968)
4. Elliott, R., Glauert, J., Kennaway, J., Marshall, I., Sáfár, E.: Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. Universal Access in the Information Society 6(4), 375–391 (2008)
5. Fernández-Soneira, A.M.: La cantidad a manos llenas: La expresión de la cuantificación en la lengua de signos española. Fundación CNSE (2008)
6. Hanke, T.: HamNoSys in a Sign Language Generation Context, pp. 249–266. Signum (2002)
7. Herrero, Á.: Escritura alfabética de la lengua de signos española: Once lecciones. Publicaciones de la Universidad de Alicante (2003)
8. Herrero, Á.: Gramática didáctica de la lengua de signos española (LSE). Ediciones SM (2009)
9. Herrero, Á., Peidro, A.: Los tipos de plural en la lengua de signos española. In: Actas del IV Congreso de Lingüística General. pp. 1051–1064 (2007)
10. Johnson, C.D.: Formal Aspects of Phonological Description. Mouton, The Hague (1972)
11. Johnston, T.: Transcription and glossing of sign language texts: Examples from Auslan (Australian Sign Language). Int. J. of Sign Linguistics 2(1), 3–28 (1991)
12. Kaplan, R.M., Kay, M.: Regular models of phonological rule systems. Computational Linguistics 20(3), 331–378 (1994)
13. Liddell, S., Johnson, R.E.: American Sign Language: The phonological base. Sign Language Studies 64, 192–277 (1989)
14. Marshall, I., Sáfár, E.: Sign language generation in an ALE HPSG. In: Müller, S. (ed.) Proc. of the HPSG-04. pp. 189–201. CSLI (2004)
15. Mohri, M., Sproat, R.: An efficient compiler for weighted rules. In: Proc. of the ACL-96. pp. 231–238 (1996)
16. San-Segundo, R., Pardo, J.M., Ferreiros, J., Sama, V., Barra-Chicote, R., Lucas, J.M., Sánchez, D., García, A.: Spoken Spanish generation from sign language. Interacting with Computers 22(2), 123–139 (2010)
17. Sandler, W.: Phonological representation of the sign: linearity and non-linearity in American Sign Language. Floris, Dordrecht (1989)
18. Schmaling, C., Hanke, T.: HamNoSys 4.0. Tech. rep. (2001)
19. Schulmeister, R.: The ViSiCAST Project: Translation into sign language generation of sign language by virtual humans (avatars) in television, WWW and face-to-face transactions. In: Stephanidis, C. (ed.) Universal Access in HCI: Towards an Information Society for All, vol. 3, p. 431–435. Erlbaum, Hillsdale, NJ (2001)
20. Shield, A., Baldridge, J.: A Morphological Analyzer for Verbal Aspect in American Sign Language, vol. 10, p. 125–138. CSLI (2009)

Modeling a Spoken Dialog System using POMDPs and Agenda-based User Simulation

David Griol, José Manuel Molina, Araceli Sanchis de Miguel[1], Zoraida Callejas[2]

[1] Dep. Computer Science, Carlos III University of Madrid, 28911 - Leganés (Spain)
{david.griol, josemanuel.molina, araceli.sanchis}@uc3m.es
[2] Dep. Languages and Computer Systems, University of Granada, CITIC-UGR, 18071 - Granada (Spain)
zoraida@ugr.es

Abstract. Spoken dialog systems in combination with mobile devices offer new opportunities in the areas of knowledge representation, natural language processing and intelligent information retrieval. In this paper, we describe the application of POMDPs and agenda-based user simulation to learn optimal dialog policies for the dialog manager in a dialog system. We have applied this approach to develop a statistical dialog manager for a dialog system which acts as a voice logbook to collect home monitored data from patients suffering from diabetes. The results of the evaluation show that the dialog system reduces the time needed to complete the dialogs with regard an initial dialog strategy, thereby allowing the dialog system to tackle new situations and generate new coherent answers for the situations already present in the initial model.

Keywords: Spoken Dialog Systems, Statistical Methodologies, POMDPs, User Simulation

1 Introduction

Speech and natural language technologies allow users to communicate in a flexible and efficient manner, making possible to access applications in which traditional input interfaces cannot be used (e.g. in-car applications, access for disabled persons, etc). Also speech-based interfaces work seamlessly with small devices and allow users to easily invoke local applications or access remote information. For this reason, spoken dialog systems [1] are becoming a strong alternative to traditional graphical interfaces which might not be appropriate for all users and/or applications.

Learning statistical approaches to model the different modules that compose a dialog system has been of growing interest during the last decade [2, 3]. The motivations for automating dialog learning are focused on the time-consuming process that hand-crafted design involves and the ever-increasing problem of dialog complexity. The most extended methodology for machine-learning of dialog strategies consists of modeling human-computer interaction as an optimization problem using Partially Observable Markov Decision Process (POMDPs) and

reinforcement methods [4]. The main drawback of this approach is due to the large state space of practical spoken dialog systems, whose representation is intractable if represented directly [5].

To solve this drawback, in this paper we use an approach that scales the POMDP framework for implementing practical spoken dialog systems by the definition of two state spaces. Approximate algorithms are also used to overcome the intractability of exact algorithms, and an alternative technique for online training based on Q-learning and an agenda-based user simulation is introduced to iteratively learn the dialog model. We have applied this proposal to develop a POMDP-based version of the rule-based Di@log system [6], which acts as a voice logbook to collect home monitored data from patients suffering from Type-2 diabetes. The comparison of this system with the initial hand-crafted dialog model shows that the POMDP methodology achieves satisfactory task completion rates and improves the selection process of goal directed actions.

The remainder of the paper is as follows. Section 2 describes the proposal to automatically learn the dialog model by means of POMDPs and the agenda-based user simulation model. Section 3 presents a detailed explanation of how this proposal has been applied to develop the Di@log system. Section 4 presents the criteria defined to carry out the evaluation of the developed dialog system and discusses the evaluation results. Finally, our conclusions and future work are presented in Section 5.

2 A dialog model with POMDPs and Agenda-based User Simulation

Formally, a Partially Observable MDP is defined as a tuple $\{S, A, T, R, O, Z, \lambda, b_0\}$ where: i) S is a set of the system states; ii) A is a set of actions that the system may take; iii) T defines a transition probability $P(s'|s, a)$; iv) R defines the immediate reward obtained from taking a particular action in a particular state $r(s, a)$; v) O is a set of possible observations that the system can receive from the world; vi) Z defines the probability of a particular observation given the state and machine action $P(o'|s', a)$; vii) λ is a geometric discount factor $0 \leq \lambda \leq 1$; and vii) b_0 is an initial belief state $b_0(s)$.

The operation of a POMDP is as follows. In each moment, the system is in an unobserved state s. The system selects an action a_m, receives a reward r, and transits to a state (unobserved) s', where s' only depends on s and a_m. The system receives an observation o' which depends on s' and a_m. Although the observation allows the system to have some evidences about the state s in which the system is now, s is not exactly known, and $b(s)$ (*belief state*) is defined to indicate the probability of the system being in the state s. Based on b, the machine selects an action $a \in A$, receives a reward $r(s, a)$, and transitions to state s', which depends only on s and a. The machine then receives an observation $o' \in O$, which is dependent on s' and a. In each moment, the probability of the system being in a specific state is updated taking into account o' and a, as shown in Equation 1.

$$b'(s') = P(s'|o', a, b) = \frac{P(o'|s'_m, a_m, b)P(s'_m|a_m, b)}{P(o'|a_m, b)} \quad (1)$$

$$= \frac{P(o'|s'_m, a_m, b) \sum_{s_m \in S_m} P(s'_m|a_m, b, s_m)P(s_m|a_m, b)}{P(o'|a_m, b)} = k \cdot P(o'|s', a) \sum_{s \in S} P(s'|a, s)b(s)$$

where $k = P(o'|a, b)$ is a normalization constant. At each time t the system receives a reward $r(b_t, a_{m,t})$ which depends on b_t and the selected action $a_{m,t}$. The reward accumulated during the dialog is called *return* and can be calculated by means of Equation 2.

$$R = \sum_{t=0}^{\infty} \lambda^t R(b_t, a_{m,t}) = \sum_{t=0}^{\infty} \lambda^t \sum_{s \in S} b_t(s) r(s, a_{m,t}) \quad (2)$$

Each action $a_{m,t}$ is determined by the policy $\pi(b_t)$ and the construction of the POMDP model implies to find the strategy π^* which maximizes the *return* at every point b. Due to the vast space of possible belief states, however, the use of POMDPs for any practical system is far from straightforward. The optimal policy can be represented by a set of policy vectors where each vector v_i is associated with an action $a(i) \in A_m$ and $v_i(s)$ equals the expected value of taking action $a(i)$ in state s. Given a complete set of policy vectors, the optimal value function and corresponding policy are computed as $V^{\pi^*}(b) = \max_i\{v_i, b\}$ and $\pi^*(b) = a(\text{argmax}_i\{v_i, b\})$ respectively.

The application of a POMDP to model a spoken dialog system is based on the classical architecture of these systems shown in Figure 1 [2]. As this figure shows, the user has an internal state S_u corresponding to a goal that is trying to accomplish and the dialog state S_d represents the previous history of the dialog. Based on the user's goal prior to each turn, the user decides some communicative action (also called intention) A_u, expressed in terms of dialog acts and corresponding to an audio signal Y_u. Then, the speech recognition and language understanding modules take the audio signal Y_u and generate the pair $(\tilde{A}_u, C)$. This pair consists of an estimate of the user's action A_u and a confidence score that provides an indication of the reliability of the recognition and semantic interpretation results. This pair is then passed to the dialog model, which is in an internal state S_m an decides what action A_m the dialog system should take. This action is also passed back to the dialog manager so that S_m may track both user and machine actions. The language generator and the text-to-speech synthesizer take A_m and generate an audio response Y_m. The user listens to Y_m and attempts to recover A_m. As a result of this process, users update their goal state S_u and their interpretation of the dialog history S_d. These steps are then repeated until the end of the dialog.

One of the main reasons to explain the challenge of building dialog systems is that $\tilde{A}_u$ usually contains recognition errors (i.e., $\tilde{A}_u \neq A_u$). As a result, the user's action A_u, the user's state S_u, and the dialog history S_d are not directly observable and can never be known to the system with certainty.

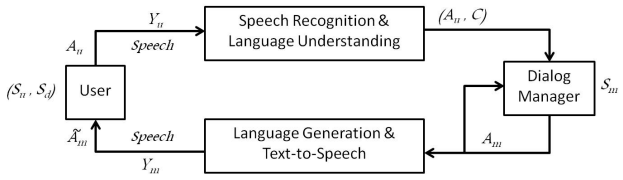


Fig. 1. Classical architecture of a dialog system

However, $\tilde{A}_u$ and C provide evidence from which A_u, S_u, and S_d can be inferred. Therefore, when using POMDPs to model a dialog system, the POMDP state S_m expresses the unobserved state of the world and can naturally be factored into three distinct components: the user's goal S_u, the user's action A_u, and the dialog history S_d. Hence, the factored POMDP state S is defined as $s_m = (s_u, a_u, s_d)$. The belief state b is then a distribution over these three components: $s_m = b_s = b(s_u, a_u, s_d)$. The observation o is the estimate of the user dialog act $\tilde{A}_u$. In the general case this will be a set of N-best hypothesized user acts, each with an associated probability $o = [(\tilde{a_u}^1, p_1), (\tilde{a_u}^2, p_2), \cdots, (\tilde{a_u}^N, p_N)]$, where $p_n = P(\tilde{a_u}^N|o)$ for $n = 1 \cdots N$.

The transition function for an SDS-POMDP follows directly by substituting the factored state into the regular POMDP transition function and making independence assumptions: $P(s'_m|s_m, a_m) = P(s'_u, a'_u, s'_d|s_u, a_u, s_d, a_m) = P(s'_u|s_u, a_m)P(a'_u|s'_u, a_m)P(s'_d|s'_u, a'_u, s_d, a_m)$. The observation model is obtained making similar reasonable independence assumptions regarding the observation function gives $P(o'|s'_m, a_m) = P(o'|s'_u, a'_u, s'_d, a_m) = P(o'|a'_u)$. The above factoring simplifies the belief update equation as shown in Equation 3.

$$b'(s'_u, a'_u, s'_d) = k \cdot \underbrace{P(o'|a'_u)}_{Observation\ model} \underbrace{P(a'_u|s'_u, a_m)}_{User\ action\ model} \sum_{s_u} \underbrace{P(s'_u|s_u, a_m)}_{User\ goal\ model} \cdot \sum_{s_d} \underbrace{P(s'_d|s'_u, a'_u, s_d, a_m)}_{Dialog\ model} b(s_u, s_d) \quad (3)$$

As shown in the previous equation, the probability distribution for a'_u is called the user action model. It allows the observation probability to be scaled by the probability that the user would speak a'_u given the goal s'_u and the last system prompt a_m. The user goal model determines the probability of the user goal switching from s_u to s'_u following the system prompt a_m. Finally, the dialog model enables information relating to the dialog history to be maintained such as grounding and focus.

Scaling the model to handle real-world problems remains a significant challenge, given that the complexity of a POMDP grows with the number of user goals, and optimization quickly becomes intractable. The Summary POMDP method [5] provides a way to scale up the POMDP model for the so-called

slot-filling spoken dialog systems. In this approach, the belief state and actions are mapped down to a summarized form where optimization becomes tractable. The original belief space and actions are called master space and master actions, while the summarized versions are called summary space and summary actions. The updated belief state b is then mapped into a summary state $\hat{b}$ where an optimized dialog policy is applied to compute a new summary machine action $\hat{a}_m$. The summary machine action is then mapped back into the master space where it is converted to a specific machine dialog act a_m.

The optimization of the policy in these two spaces is usually carried out by using techniques like the Point-based Value Iteration or Q-learning, in combination with a user simulator. Q-learning is a technique for online learning where a sequence of sample dialogs are used to estimate the Q functions for each state and action. The optimal action for each point p is given by $\bar{a}_p = \text{argmax}_{\bar{a}}\, \bar{Q}(a, p)$. After a set of dialogs has been completed, the estimates of the Q-functions are updated with the new dialog scores. Simulation is usually done at a semantic dialog act level to avoid having to reproduce the variety of user utterances at the word or acoustic levels. At the semantic level, at any time t, the user is in a state s_u, takes action a_u, transitions into the intermediate state s'_u, receives machine action a_m, and transitions into the next state s''_u. Agenda-Based state representations, like the one described in [7], factors the user state into an agenda A and a goal G. The goal G consists of constraints C which specify the detailed venue of the dialog, and requests R which specify the desired pieces of information.

The user agenda A is a stack-like structure containing the pending user dialog acts that are needed to elicit the information specified in the goal. At the beginning of each dialog, a new goal G is randomly selected. Then, the goal constraints C are converted into user and system *inform acts* (a_u and a_m acts) and the requests R into *request acts*. A *bye* act is added at the bottom of the agenda to close the dialog once the goal has been fulfilled. The agenda is ordered according to priority, with $A[N]$ denoting the top and $A[1]$ denoting the bottom item. In response to incoming machine acts a_m, new user acts are pushed onto the agenda and no longer relevant ones are removed. This way, the agenda tracks the progress of the dialog and encodes the complete dialog history. $a_u[i]$ denotes the *ith*-item in the user act a_u, which is calculated by means of Equation 4.

$$\begin{aligned} a_u[i] &:= A[N-n+i] \quad \forall i \in [1..n];\ \ 1 \leq n \leq N \\ P(a_u|S) &= P(a_u|A,G) = \delta(a_u, A[N-n+1..N]) \end{aligned} \tag{4}$$

By denoting A' to the agenda after selecting the act a_u, the first state transition depending on a_u can be expressed by means of Equation 5.

$$P(S'|a_u, S) = P(A', G'|A, G) = \delta(A', A[1..N'])\delta(G', G) \tag{5}$$

Using $S = (A, G)$, the second state transition based on a_m can be decomposed into two terms respectively related to the goal update and the agenda update modules, as shown in Equation 6.

$$P(S''|a_m, S') = P(a_m|A', G'')P(G''|a_m, G') \tag{6}$$

By assuming that every dialog act a_m triggers one push-operation from the agenda, the probability related to the agenda update model can be expressed as shown in Equation 7.

$$P(a_m|A', G'') == \prod_{M}^{i=1} P(A''[N'+i]|a_m[i], G'') \;\; \delta(A''[1..N'], A'[1..N']) \tag{7}$$

Assuming that R'' is conditionally independent of C' given C'', the second probability denoting the goal update model can be expressed as shown in Equation 8.

$$\begin{aligned} P(G''|a_m, G') &= P(R''|a_m, R', C'')P(C''|a_m, R', C') \\ P(R''|a_m, R', C'') &= \prod_{k} P(R''[k]|a_m, R'[k], \mathcal{M}(a_m, C'')) \end{aligned} \tag{8}$$

To simplify $P(C''|a_m, R', C')$, it is assumed that C'' is derived from C'. This way, the choice of a transition does not need to be conditioned on the full space of possible a_m, R' and C'.

3 Practical Application: The DI@L-log dialog system

In this paper we focus on the use of POMDP's and agenda-based simulation to model the user and dialog systems. As a practical application we have selected DI@L-log, a dialog system which acts as a voice logbook to collect home monitored data from patients suffering from Type-2 diabetes [6]. The data collected by the system are the patient's weight, blood pressure (systolic and diastolic values) and sugar levels. The system validates and analyzes the data, providing some immediate feedback to the patients regarding their current progress as well as communicating the results to doctors at the hospital who are able to review the patient's progress graphically and deal with any alerts generated by the system concerning abnormal developments.

To apply our methodology to this task, the definition of the semantics was carried out considering the information that is required to monitor the patients and inform them about their condition. Taking into account the requirements of the task, an initial strategy was designed for the DM. Figure 2 shows an example of a dialog acquired using this strategy. As can be observed, three different phases are present in every dialog. Firstly, there is an identification phase in which the system asks the user about his login and password and then waits until the user says that he is ready to provide the control data (S1 and S2 system turns). Secondly, the system analyzes which data is required for the current user, taking into account that the weight and sugar values are mandatory and the blood

control is only carried out for specific patients (S3 to S10 system turns). Every item is confirmed after the user has provided its value and the user can only provide one item at a time. In the last phase, the system checks the information that the patient has provided during the current dialog and compares it with the data that is present in a database that contains the values that he provided in previous dialogs. By means of this comparison, the system is able to inform the user about his condition and provide warnings and hints.

S1: Welcome to the DI@L-log System. Please say or enter your four digit pin.
U1: 1356.
S2: Are you ready to give your measurements?
U2: Yes.
S3: Let's begin with your weight. How many stones?
U3: 12.
S4: I have understood 12 stones. Is it correct?
U4: Yes.
S5: What is your blood sugar?
U5: 11.
S6: I have understood 11. Is it correct?
U6: Yes.
S7: And finally, for your blood pressure you will be asked for the top number and then the lower number. What is the top number of your blood pressure?
U7: 160.
S8: I have understood 160. Is it correct?
U8: Yes.
S9: And what is the lower number?
U9: 86.
S10: I have understood 86. Is it correct?
U10: Yes.
S11: Your readings have been successfully updated. According to them, your weight is higher than last week, be very careful. Thanks for contacting DI@L-log. Goodbye.

Fig. 2. An example of a dialog for the DI@L-log task

A corpus of 100 dialogs was acquired using this strategy. In order to learn statistical models, the dialogs of the corpus were labeled in terms of dialog acts. In the case of user turns, the dialog acts correspond to the classical frame representation of the meaning of the utterance. A frame is a structure for representing a concept or situation. Each concept in a domain has usually associated a group of attributes (slots) and values. For the DI@L-log task, we defined three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*) and four attributes (*Weight*, *Sugar*, *Systolic-Pressure*, and *Diastolic-Pressure*).

The labeling of the system turns is similar to the labeling defined for the user turns. A total of 12 task-dependent concepts was defined, corresponding to the set of concepts used by the system to acquire each of the user variables (*Weight*, *Sugar*, *Systolic-Pressure*, and *Diastolic-Pressure*), concepts used to confirm the values provided by the user (*Confirmation-Weight*, *Confirmation-Sugar*, *Confirmation-Systolic*, and *Confirmation-Diastolic*), concepts used to inform the patient about his condition (*Inform*), and three task-independent concepts (*Not-Understood*, *Opening*, and *Closing*).

4 Results of the Evaluation

The summary Q-learning algorithm and agenda-based user simulation described in Section 2 were used to develop a POMDP-based dialog system for the Di@L-log task. Rewards in the dialog system were based on the task completion rate and the number of turns in the dialog. The agenda-based user simulator was initially trained using the 100 dialogs acquired with the initial strategy previously described. A POMDP-based dialog manager was implemented and trained via interactions with the simulated user model to iteratively learn a dialog policy. A total of 150,000 dialogs were simulated. Using the definitions described in [7], the POMDP system obtained 20 points for a successful dialog and 0 for an unsuccessful one. One point was subtracted for each dialog turn. System evaluation was then carried out through the validation of the simulation experiments by means of a study with real users. A total of 100 new dialogs was recorded from interactions of ten users employing the dialog policy learned using our proposal with the same noise conditions of the initial acquisition.

The first group of experiments to compare the initial dialogs with the dialogs acquired with the proposed POMDP-based methodology covered the following statistical properties to evaluate the quality of the dialogs obtained using different dialog strategies: i) Dialog length, measured as the number of turns per task; number of turns of the shortest dialog; number of turns of the longest dialog; and number of turns of the most seen dialog; ii) Different dialogs in each corpus, measured as the percentage of different dialogs (different labeling and/or order of dialog acts) and the number of repetitions of the most observed dialog; iii) Turn length, measured as the number of actions per turn; and iv) Participant activity, measured as the ratio between system and user actions per dialog. Table 1 shows the comparison of the different high-level measures for the initial corpus and the corpus acquired with the POMDP-based methodology.

	Initial Dialog System	**Final Dialog System**
Average number of turns per dialog	12.9±2.3	7.4±1.6
Percentage of different dialogs	22.9%	48.3%
Repetitions of most seen dialog	38	3
User turns of most seen dialog	9	7
User turns of shortest dialog	7	5
User turns of longest dialog	13	9

Table 1. Results of the high-level dialog features defined for the comparison of the dialogs for the initial and final dialog systems

The first improvement that can be observed is a reduction in the number of turns. This reduction can also be observed in the number of turns of the longest, shortest and most seen dialogs. These results show that improving the dialog strategy makes it possible to reduce the number of necessary system actions. The greater variability of the resulting dialogs can be observed in the higher percentage of different dialogs and less repetitions of the most seen dialog

obtained with the final dialog strategy. There is also a slight increment in the mean values of the turn length measured as the number of user actions per turn. The dialogs acquired with the final strategy are statistically longer, as they showed 1.6 actions per user turn instead of the 1.3 actions observed in the initial dialogs. This is also due to the better selection of the system actions.

The second group of experiments covered the following statistical properties: frequency of different user and system actions (dialog acts), and proportion of goal-directed actions (request and provide information) versus grounding actions (confirmations). We consider as well the remaining possible actions. In both cases, significant differences in the dialog acts distribution were observed. With regard to user actions, users need to employ less confirmation turns in the final strategy (from 14.7% to 11.3%), which explains the higher proportion for the rest of user actions in this strategy. It also explains the lower proportion of yes/no actions in the final strategy (from 18.6% to 14.5%), which are mainly used to confirm that the system's service has been correctly provided. With regard to the system actions, there is a reduction in the number of system requests for data items (from 30.1% to 19.8%). This explains a higher proportion of turns to inform and confirm data items in the dialogs of the final strategy. In addition, we grouped user and system actions into categories in order to compare turns to request and provide information (goal directed actions) versus turns to confirm data items and make other actions (grounding actions). This study also showed the better quality of the dialogs in the final strategy, given that the proportion of goal-directed actions is higher in these dialogs (from 56.6% to 69.2%).

Finally, Figure 3 shows the evolution of the average duration of the dialogs during the simulation process in terms of total dialog turns (*duration*). It also shows the reduction in the number of responses provided by the dialog system which cause a failure of the dialog (*#error*) and the number of unseen situations for which there is not a system response in the dialog model (*#unseen*). As it can be seen from these results, the final dialog system not only reduces the time required to fulfill each one of the objectives of the dialog, but also it reduces the possibility of selecting an erroneous response. Therefore, the enhanced dialog policy allows to tackle new situations and generate new coherent answers for the situations already present in the initial corpus and automatically detect different valid paths to achieve each of the required objectives and reduce the number of errors.

5 Conclusions

POMDPs and reinforcement methods are the most extended methodology for modeling dialog systems. However, this approach has the drawback that the state space of practical dialog systems can become intractable if represented directly. In this paper, we have explored a proposal to deal with this problem based on the creation of a summarized belief space for states and actions and the introduction of an agenda-based user simulation to learn optimal dialog policies. The application of this approach to the Di@log system shows how it allows the

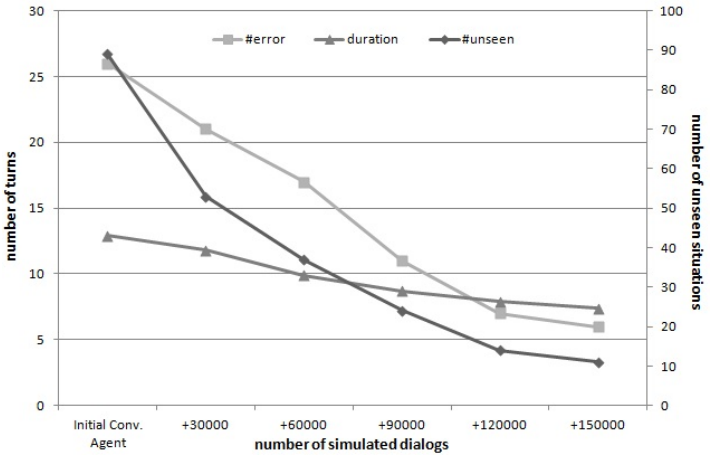


Fig. 3. Evolution of the number of unseen situations, number of errors, and average number of turns

dialog manager to tackle new situations and generate new coherent answers for the situations already present in an initial strategy. This way, the dialog system can automatically detect different valid paths to achieve the required objectives, confirm information items taking into account the confidence scores, and reduce the duration of the dialogs. As a future work, we are adapting our proposal for its application in other domains. We also want to compare this approach with other statistical methodologies for dialog management.

References

1. López-Cózar, R., Araki, M.: Spoken, Multilingual and Multimodal Dialogue Systems. John Wiley & Sons Publishers (2005)
2. Young, S.: The Statistical Approach to the Design of Spoken Dialogue Systems. Technical report, TR.433, Cambridge University Engineering Department (2002)
3. Griol, D., Hurtado, L., Segarra, E., Sanchis, E.: A Statistical Approach to Spoken Dialog Systems Design and Evaluation. Speech Communication **50**(8–9) (2008) 666–682
4. Levin, E., Pieraccini, R., Eckert, W.: A stochastic model of human-machine interaction for learning dialog strategies. IEEE Transactions on Speech and Audio Processing **8**(1) (2000) 11–23
5. Young, S., Schatzmann, J., Weilhammer, K., Ye, H.: The Hidden Information State Approach to Dialogue Management. In: Proc. of ICASSP'07. (2007) 149–152
6. Black, L., McTear, M., Black, N., Harper, R., Lemon, M.: Appraisal of a conversational artefact and its utility in remote patient monitoring. In: Proc. of CBMS'05. (2005) 506–508
7. Thomson, B., Schatzmann, J., Weilhammer, K., Ye, H., Young, S.: Training a real-world POMDP-based Dialog System. In: Proc. of HLT'07. (2007) 9–16

Methodology for developing a Speech into Sign Language Translation System in a New Semantic Domain

Verónica López-Ludeña[1], Rubén San-Segundo[1], Carlos González[2], Juan Carlos López[2] y José M. Pardo[1]

[1]Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica. ETSI Telecomunicación. Universidad Politécnica de Madrid
[2]Grupo de Sistemas Inteligentes Aplicados, Departamento de Tecnologías y Sistemas de Información. Universidad de Castilla-La Mancha

veronicalopez@die.upm.es

Abstract. This paper proposes a methodology for developing a speech into sign language translation system considering a user-centered strategy. This methodology consists of four main steps: analysis of technical and user requirements, data collection, technology adaptation to the new domain, and finally, evaluation of the system. The two most demanding tasks are the sign generation and the translation rules generation. Many other aspects can be updated automatically from a parallel corpus that includes sentences (in Spanish and LSE: Lengua de Signos Española) related to the application domain. In this paper, we explain how to apply this methodology in order to develop two translation systems in two specific domains: bus transport information and hotel reception.

Keywords: Methodology, New domain, Deaf people, Spanish Sign Language (LSE), Spoken Language Translation, Sign Animation

1 Introduction

Defining a methodology is an important aspect in order to develop human-computer interaction systems. A methodology ensures that quality requirements are reached (performance, time and cost). The usability of the developed system will depend on that quality and, as a result of a higher usability, there will be a better acceptance from deaf people. In this paper, we explain the methodology used in the CONSIGNOS project (Plan Avanza Exp Nº: TSI-020100-2010-489). This project aims to adapt a Spanish into Spanish Sign Language (LSE: Lengua de Signos Española) translation system, already developed in other specific domains (like driver's license renewing service), to different domains: bus transport information and hotel reception.

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

2 State of the art

Two of the most relevant projects in speech into sign language translation, ViSiCAST and eSIGN (Elliot et al., 2008), follow a specific methodology. The ViSiCAST project (Elliot et al., 2008), whose main goal was to produce adaptable communication tools allowing sign language communication, was structured to have three application-oriented work packages, each focusing on the technical issues in delivery for that specific application area, and two enabling technology work packages, focusing on virtual signing, sign language representation, and sign language synthesis from conventional textual sources. A further evaluation work package is concerned with eliciting feedback from deaf people at various stages within the development of the system.

The eSIGN project aimed to provide sign language on websites. The different tasks of this project are: development of tools needed for creating signed content; improvement of the signed output the avatar; creating the first information sites on the Internet with animated signing; content creation in all three partner countries; the further development of tools needed for creating signed content; further improvement of the signed output of the avatar and the user involvement and continued evaluation of their tools and the avatar's comprehensibility.

A similar Spanish project "Speech into Spanish Sign Language Translation for a personal service" (San Segundo et al., 2008; San Segundo et al., 2011) had the following developing tasks: linguistic study, system design, speech recognizer design, rule-based translation development, statistic translation development, sign generation, evaluation and documentation of the project.

Another example of sign language translation systems are: VANESSA (Voice Activated Network Enabled Speech to Sign Assistant) project (Tryggvason, 2004) that was part of eSIGN and that facilitates the communication between assistants and their deaf clients in UK Council Information Centres (CIC's) or similar environments. Other two main research projects that focus on sign language recognition are DICTA-SIGN (Hanke et al., 2010; Efthimiou et al., 2010) and SIGN-SPEAK (Dreuw et al., 2010a and 2010b). DICTA-SIGN aims to develop the technologies necessary to make Web 2.0 interactions in sign language possible. In SIGN-SPEAK, the overall goal is to develop a new vision-based technology for recognizing and translating continuous sign language into text.

There are many other examples of design, development and evaluation of Sign Language translation systems. In (Cox et al., 2003), it is described the design, implementation and testing of an experimental interactive translation system that aims to aid a deaf person in transactions in a Post Office and they also explain the evaluation of the system. The paper (Huenerfauth et al., 2008) describes an implementation and user-based evaluation (by native ASL signers) of a prototype ASL natural language generation system that produces sentences containing classifier predicates, which are frequent and complex spatial phenomena that previous ASL generators have not produced.

3 Methodology overview

This paper presents an adaptation of the Participatory Design methodology: one of the most used User-Centered Design approaches that follows the ISO standard Human-centered design for interactive systems: ISO 9241-210, 2010. Participatory design (known before as 'Cooperative Design') is a designing approach where all stakeholders (e.g. employees, partners, customers, citizens, and end-users) are involved actively in the design process. The main target is to guarantee that the final designed product meets their needs and is usable. This methodology consists of the following phases or tasks (**Fig. 1**):

- The requirement analysis (phase 1) will be undertaken with two Participatory Design workshops where end-users (Deaf people), researchers and developers will work together to define the system and user requirements.
- The data collection (phase 2) will be performed using preliminary prototypes.
- During technology adaptation (phase 3), users must supervise the process in order to guarantee that all user requirements are considered.
- The evaluation (phase 4): evaluation design, field evaluation and results analysis.

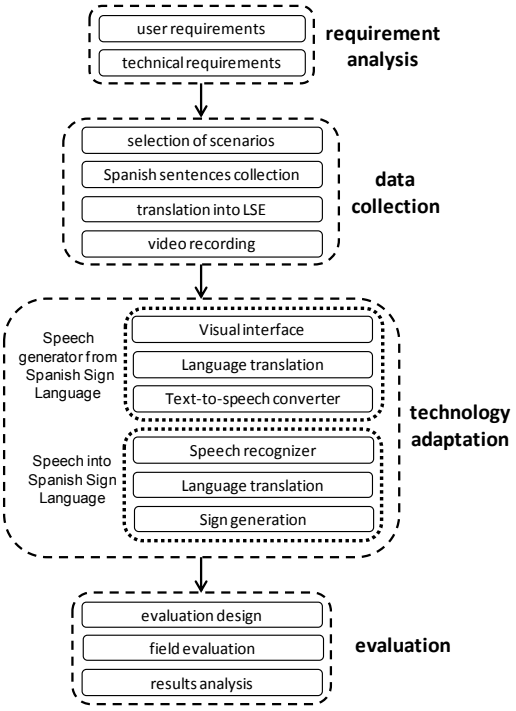


Fig. 1. Phases of system adaptation to a new application domain

The following sections explain the application of each phase.

4 Requirement analysis (Phase 1)

This section describes the process to analyze and to collect the main requirements of deaf users in relation to the development of an advanced communication system that will facilitate the communication between deaf and hearing people.

4.1 User requirements

According to the Survey of Disability, Personal Autonomy and Dependency Situations (EDAD, 2008) from INE (Spanish Statistic Institute), there are 1,064,100 deaf people in Spain. Deafness brings about significant communication problems: most deaf people have problems when expressing themselves in oral languages or understanding written texts. Their communication barriers are the main cause of the fact that 47% of deaf population has not studies or even is illiterate (INE -Spanish Statistic Institute- 1999 y MEC -Science and Education Ministry- 2000/2001).

All these aspects support the need to generate new technologies in order to develop automatic translation systems for converting this information into sign language.

First of all, it is necessary to define the scenarios of the system that we need to develop. In the case of our translation system, two different scenarios have been considered that correspond to personal attention services. In these two scenarios, two independent demos will be developed: the first scenario is the personal attention service that the EMT (Municipal Transport Enterprise) provides to users. The second is the personal attention service in a hotel reception. For each one of these services, the next steps have been followed for the analysis and definition of requirements:

- Visit to the facilities of the personal attention service site and interviews with employees. 3-4 interviews were carried out with the employees of EMT and the hotel in order to define the application domain for the demos.
- General description of service; the most requested information by users; service schedule and distribution of consultations per schedules; description of information offered to users; kind of users, etc.

4.2 Technical requirements

An important challenge of the project is to achieve a minimum level of quality of the technology, because the usability of the developed systems will depend on that quality. The following technologic requirements are defined:

- Speech recognition system will demand a recognition rate higher than 90% in the selected application domain. If that rate is not reached with speaker-independent models, an adaptation process will be performed to each speaker involved in the evaluation in order to guarantee that rate.
- It is also necessary a translation rate higher than 90% for the application domains targeted in the project. In order to obtain that translation rate, several translation strategies will be analyzed and combined.

– Finally, it is necessary an intelligibility of the avatar of 90% when representing the signs: recognition rate of deaf people. In order to obtain this intelligibility, the sign generation uses techniques based on inverse kinematics and semi-automatic movement capture that lets to obtain more realistic movements. This approximation requires more time for the dictionary generation, but it is more realistic.

In order to guarantee these technological requirements, it will be required a Spanish-LSE parallel corpus with a significant number of sentences in the application context. An approximated estimation for each service would be around 500 sentences containing less than 1,000 Spanish words and less than 200 signs in Spanish Sign Language.

5 Data collection (Phase 2)

5.1 Sentence recording process

In order to obtain the database, the next four steps are necessary:

– Selection of scenarios and application domain. The application domains were selected based on a guided visit to the installations and a set of interviews with the employees that offer the services
– Collection of sentences in Spanish. Once we defined the scenarios or application domains, researchers went to different point of service and collected sentences that were pronounced by both users and employees. This collection took several days.
– Translation of Spanish sentences into LSE (Spanish Sign Language). All Spanish sentences were translated by Deaf people that know Spanish language.
– Video recording. All sentences were represented by LSE experts.

5.2 Database statistics

	EMT		**HOTEL**	
	Spanish	**LSE**	**Spanish**	**LSE**
Total number of pair of sentences	600	600	499	499
Different sentences	523	362	493	333
Total number of words or glosses	3801	2800	3850	2780
Vocabulary	627	303	734	350

Table 1. Main statistics of the corpus

Table 1 shows the main characteristics of the corpus. The "vocabulary" row indicates the number of different words or glosses. There are 523 different sentences from the EMT information service and 493 different sentences from different situations in a

hotel reception, with corresponding translation into LSE. There are fewer different LSE sentences than Spanish sentences, because different Spanish sentences can have the same LSE translation.

6 Adapting the speech and language technologies to a new domain (Phase 3)

This section describes the main aspects that must be considered when increasing the system scope or when applying this system to other domain.

6.1 Speech into Spanish Sign Language translation system

The Spanish into LSE translation system converts natural speech sentences into LSE sentences signed by an avatar. This system is made up of three modules (**Fig. 2**). The first one is a speech recognizer that converts natural speech into a sequence of words (text). The second one is a natural language translation module that converts a word sequence into a sign sequence. And the third one is an avatar that represents the signs.

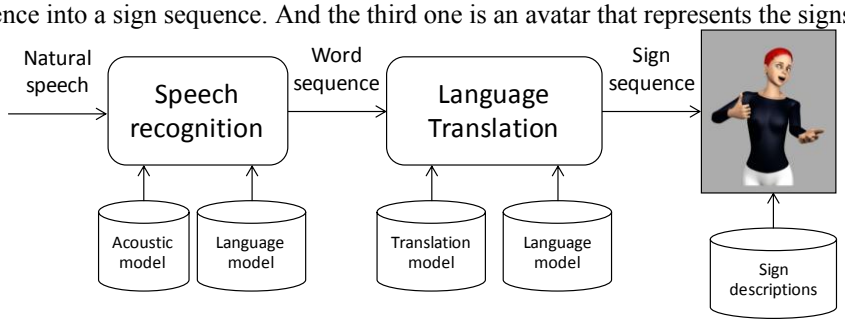


Fig. 2. Speech into Spanish Sign Language translation system

In order to adapt this system, it is necessary update the three modules.

- **Speech recognizer**

Vocabulary and language models of the speech recognition system must be updated using the Spanish sentences collected for each application domain. The speech recognizer includes an acoustic adaptation module for adapting the acoustic models to a new specific environment, a new speaker, or a new Spanish accent. Also, when generating automatically the vocabulary and a language model for the speech recognizer, a new module has been included for introducing source language variants, increasing the speech recognizer flexibility.

- **Language translation module**

The translation module has a hierarchical structure divided into two main steps. In the first step, an example-based strategy is used to translate the word sequence. And there

is also a background module that translates the sentence if the previous translation is not good enough. For the background module, a combination of rule-based (where a set of translation rules, defined by an expert, guides the translation process) and statistical translators (phrase-based translator and Finite State Transducers) has been used.

For the example-based translation module, it is necessary to update the examples of the database. These examples consist of Spanish sentences and their corresponding translation (the parallel corpus described in section 5).

The rule-based translation module would need to develop new rules for translating new sentences. It is a time-consuming task because an expert must develop the rules by hand. Some of the rules (approx. 40%) are general translation rules and they can be used in other domains, but a lot of them are specific for this domain. The proposed rules used in the system were developed by one person during 3 weeks.

For the statistical translation, it is necessary to update the translation models: these models are obtained automatically from a parallel corpus (section 5). Because the rule-based strategy is the most demanding task, the statistical translation strategy incorporates a new pre-processing module (López-Ludeña et al, 2011) that permits to increase its performance, and to replace the rule-based translation strategy.

- **Sign representation**

The animation module uses a declarative abstraction module used by all of the internal components. This module uses a description based on XML, where each key pose configuration is stored defining its position, rotation, length and hierarchical structure. We have used an approximation of the standard defined by H-Anim (Humanoid Working Group ISO/IEC FCD 19774:200x). In terms of the bones hierarchy, each animation chain is composed by several « joint » objects that define transformations from the root of the hierarchy.

In order to increase its adaptability, the sign animation module includes a new version of the sign editor that incorporates new options (like predefined positions and orientations) for reducing significantly the sign specification time.

6.2 Spanish generator from Spanish Sign Language

For the Spanish generator from LSE (**Fig. 3**), the necessary changes affect two of the three modules composing the system: visual interface and language translation. The text-to-speech conversion module works well for any sentence in Spanish, so it is not necessary to introduce any change on it.

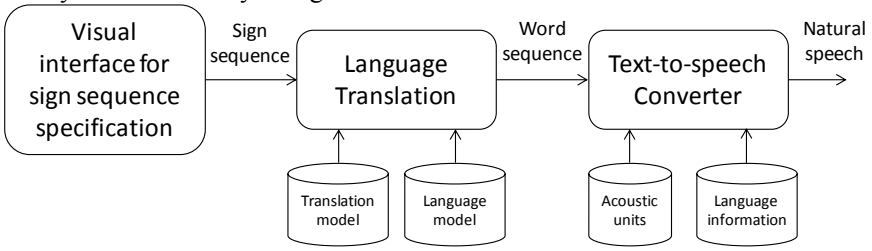


Fig. 3. Spanish generator from Spanish Sign Language

- **Visual interface.**

It is necessary to update the list of glosses considered in the search. When a deaf user specifies a certain gloss, an avatar represents it in order to allow deaf user checking that the selected gloss is appropriate. In order to add a new sign to the interface, the system needs to have a new file in a specific path, named with the sign gloss: e.g. CAR.txt. This file contains the sign description, mandatory to be represented by the avatar. This description can be generated using a sign Editor. When a new file is detected in the path, the interface updates the search with a new gloss (file name) and the avatar can represent it.

Generating the signs is the most time-consuming task, because every sign file must be generated by hand. For example, for generating 715 signs (needed in these two domains), it was necessary one person working during 2 month. It is true that when one sign has been generated, it can be reused in any domain. On the other hand, the time and resources for generating a sign are lower compared to the option of recording a video for every sign. A new Editor has been developed reducing the developing time more than 50% approx. by using techniques based on inverse kinematics and semi-automatic movement capture that lets to obtain more realistic movements.

- **Translation module**

It is the same module as commented in the previous section, but in opposite translation direction. So, the update is the same.

7 Evaluation (Phase 4)

7.1 Design and field evaluation

In order to carry out a system evaluation it is necessary to take into account several aspects. The evaluation should be carried out in at least two days and around twelve people should be involved: 2 employees of each application domain and 10 deaf users. On the other hand, it is necessary to define different scenarios (around five or six) in order to test real situations. For instance, asking for different kinds of information or service or simulating different kind of problems. The ten users should interact with the two employees and all of them should test the system in almost all the scenarios. This way, finally we should have around 50 dialogues.

It is important to collect the age's range of deaf users, their understanding level of written Spanish, their habit of using glosses for sign sequence specification and their practice with computers.

7.2 Objective measurements

Evaluation must include objective measurements from the system and subjective information from both user and employee questionnaires.

The objective measurements can be obtained using a capturing software (Camtasia Studio 6: http://www.techsmith.com/camtasia.html) and a detailed log generated by the system. These measurements should include the next aspects: speech recognition rate, translation rate, average translation time, average time for text to speech conversion, percentage of cases using example-based translation, percentage of cases using each translation technique (example-based, rules-based or statistical), time for gloss sequence specification, number of user turn per dialogue, number of dialogues, etc.

7.3 Subjective measurements

Subjective measurements can be collected from questionnaires filled by both employees and deaf users. In these questionnaires, they could evaluate different aspects of the system giving a score between 0 and 5. Designing questionnaires for Deaf is an interesting aspect with several problems that it is important to deal with. A group of experts have to design the questionnaires taking into account the following aspects:

- To decide the language for asking the different questions: LSE (using videos) or written Spanish. A good solution would be to present the questions in Spanish with translation in LSE (glosses) and having two interpreters for solving any question.
- To decide the aspects to evaluate and design questions. Experts in LSE (Deaf) report problems with concepts/words used in questionnaires developed for evaluating Speech-based applications (Möllera et al, 2007) or Human-Computer Interaction (HCI) systems (Brooke, 1996), because these questionnaires have not translation into LSE, so many of these concepts would be difficult to understand by Deaf). Due to this aspect, experts advise to reduce the number of questions.
- Finally, the scale used: number of levels and the names for the different levels. For the number of levels, the expert panel advise to define an even number (six in this case) eliminating the neutral level and forcing the user to decide. One reason is because this neutral level is the most common refuge when a user does not understand one of the questions. Forcing a user to decide provokes that this user asks more questions to the interpreter in order to understand all the details.

8 Conclusions

This paper has described a methodology for developing a speech into sign language translation system considering a user-centered strategy. This methodology consists of four main steps: analysis of technical and user requirements, data collection, technology adaptation to the new domain and, finally, evaluation of the system. In this paper we have explained how to apply this methodology in order to develop two translation systems in two specific domains: bus information and hotel reception.

It is possible to conclude that many aspects can be updated automatically from the generated parallel corpus. Except sign specifications and translation rules that are the two most demanding tasks. In the case of sign generation, it is useful a new editor for sign specification that incorporates new options for reducing significantly the sign specification time. In the case of translation rules, the authors are considering to work

over the statistical translation module in order to increase its performance and to delete the rule-based method.

9 Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement n° 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542), CONSIGNOS (Plan Avanza Exp Nº: TSI-020100-2010-489) and the European FEDER fund projects.

10 References

1. Brooke, J.: SUS: *A 'quick and dirty' usability scale*. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London, 1996.
2. Cox, S.; Lincoln, M.; Nakisa, M.; Wells, M.; Tutt, M. & Abbott, S. The Development and Evaluation of a Speech-to-Sign Translation System to Assist Transactions International Journal of Human-Computer Interaction, 2003, 16, 141-161
3. Dreuw P., Ney H., Martinez G., Crasborn O., Piater J., Miguel Moya J., and Wheatley M., 2010. *The Sign-Speak Project - Bridging the Gap Between Signers and Speakers*. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, 2010a. pp 73-80.
4. Dreuw P., Forster J., Gweth Y., Stein D., Ney H., Mar-tinez G., Verges Llahi J., Crasborn O., Ormel E., Du W., Hoyoux T., Piater J., Moya Lazaro JM, and Wheatley M. 2010. *SignSpeak - Understanding, Recognition, and Translation of Sign Languages*. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, May 2010b. pp 65-73.
5. Efthimiou E., Fotinea S., Hanke T., Glauert J., Bowden R., Braffort A., Collet C., Maragos P., Goudenove F. 2010. *DICTA-SIGN: Sign Language Recognition, Generation and Modelling with application in Deaf Communication*. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, May 2010. pp 80-84.
6. Elliott, R. and Glauert, J.R.W. and Kennaway, JR and Marshall, I. and Safar, E. *Linguistic modelling and language-processing technologies for Avatar-based sign language presentation*. Universal Access in the Information Society, Vol. 6, No. 4, pp 375-391, Springer, 2008.
7. Hanke T., König L., Wagner S., Matthes S. 2010. *DGS Corpus & Dicta-Sign: The Hamburg Studio Setup*. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, May 2010. pp 106-110.
8. Huenerfauth, M.; Zhao, L.; Gu, E. & Allbeck, J. *Evaluation of American Sign Language Generation by Native ASL Signers* ACM Transactions on Accessible Computing, ACM, 2008, 1, 1-27
9. López-Ludeña, V., San-Segundo, R., Montero, J.M., Córdoba, R., Ferreiros, J., Pardo, J.M., *Automatic Categorization for Improving Spanish into Spanish Sign Language Ma-*

chine Translation. In press Computer Speech and Language, ISSN: 0885-2308, Available online 1 October 2011

10. Möllera, S., Smeeleb, P., Bolandc H., and Krebbera, J.,. *Evaluating spoken dialogue systems according to de-facto standards: A case study*. Computer Speech & Language. Volume 21, Issue 1, January 2007, Pages 26-53.
11. San-Segundo R., Barra R., Córdoba R., D'Haro L.F., Fernández F., Ferreiros J., Lucas J.M., Macías-Guarasa J., Montero J.M., Pardo J.M, 2008. *Speech to Sign Language translation system for Spanish*. Speech Communication, Vol 50. 2008. pp. 1009-1020.
12. San-Segundo, R., Montero, J.M., Córdoba, R., Sama, V., Fernández, F., D'Haro, L.F., López-Ludeña, V., Sánchez D., and García, A., 2011. *Design, development and field evaluation of a Spanish into sign language translation system*. Pattern Analysis and Applications. To appear.
13. Tryggvason, Judy., 2004. *VANESSA: A System for Council Information Centre Assistants to communicate using sign language"* School of Computing Science de la Universidad de East Anglia.

Architecture for Text Normalization using Statistical Machine Translation techniques

V. López-Ludeña, R. San-Segundo, J. M. Montero, R. Barra-Chicote, J. Lorenzo

Speech Technology Group. E.T.S.I. Telecomunicación.
Universidad Politécnica de Madrid. Spain.

veronicalopez@die.upm.es

Abstract. This paper proposes an architecture, based on statistical machine translation, for developing the text normalization module of a text to speech conversion system. The main target is to generate a language independent text normalization module, based on data and flexible enough to deal with all situations presented in this task. The proposed architecture is composed by three main modules: a tokenizer module for splitting the text input into a token graph (tokenization), a phrase-based translation module (token translation) and a post-processing module for removing some tokens. This paper presents initial experiments for numbers and abbreviations. The very good results obtained validate the proposed architecture.

Keywords: text normalization, text to speech conversion, language translation, numbers, acronyms, abbreviations.

1 Introduction

Although Text to Speech (TTS) conversion is the area where more effort is devoted to text normalization, dealing with real text is a problem that also appears in other applications like machine translation, topic detection and speech recognition. In an ideal situation, there would be an unambiguous relationship between spelling and pronunciation. But in real text, there are not ordinary words like numbers, digit sequences, acronyms, abbreviations, dates, etc.

The main problem of a text normalization module consists of converting Non-Standard Words (NSWs) into regular words. This problem can be seen as a translation problem between a real text (including NSWs) and an ideal text where all the words are standard: there is a unique relationship between word spelling and its pronunciation.

2 State of the art

One of the main references focused on text normalization is (Sproat et al, 2001). In this reference, authors propose a very complete taxonomy of NSWs considering 23

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

different classes grouped in three main types: numerical, alphabetical and miscellanea. Sproat et al describe the whole normalization problem of NSWs, proposing solutions for some of the problems: a good strategy for tokenizing the input text, a classifier for determining the class associated to every token, some algorithms for expanding numeric and other classes that can be handled "algorithmically", and finally, supervised and unsupervised methods for designing domain-dependent abbreviation expansion modules.

Additionally, it is also important to remark other references that have addressed specific problems included in the text normalization research line. Focused on abbreviations and acronyms, there are several efforts focused on extracting them from text automatically (Yeates, 1999; Larkey et al., 2000; Chang et al., 2002) and other efforts trying to model how they are generated (Cannon, 1989; Pennell et al., 2011a). Numbers (Sproat, 2010) and proper names (Bikel et al. 1999; Collins et al., 1999; Jonnalagadda and Topham, 2010) have been also the target of other research works.

Nowadays, much effort on text normalization is focused on SMS language, interchanged through mobile phones and social networks like Facebook or Twitter (Brody et al., 2011, Han et al., 2011 and Kaufmann, 2010).

Due to the important advances obtained in machine translation in the last decade, there has been an increasing interest on exploring the machine translation capabilities for dealing with the problem of text normalization (Aw et al. 2006; Pennell and Liu, 2011b).

This paper proposes a general architecture based on statistical machine translation techniques for text normalization. The main target is to generate a language independent text normalization module, based on data (instead of on expert rules) and flexible enough to deal with all situations. This paper presents initial experiments for numbers and abbreviations for Spanish.

3 Architecture Description

Fig. 1 shows the architecture diagram of the text normalization module proposed in this paper.

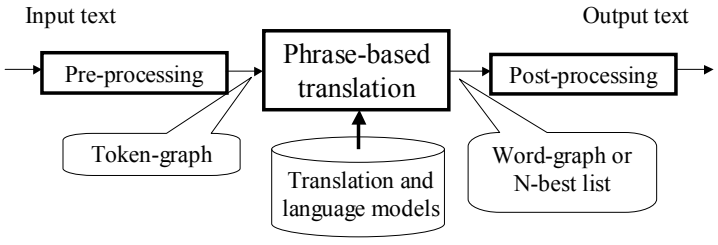


Fig. 1. Architecture diagram

This architecture is composed by three modules: a preprocessing module that splits the text input into a token graph (tokenization), a phrase-based translation module (token translation) and a post-processing module for removing some tokens.

3.1 Preprocessing: sentence tokenization

At this first module, the input text is split into tokens. This process is carried out in two different steps.

At the first step, a preliminary token sequence is generated considering a small set of rules. As one of the main targets of this work is to provide a language independent architecture, the main rules should be language independent:

- The first rule supposes that blank characters provide an initial segmentation in tokens.
- The second rule subdivides initial tokens (sequence of characters between blank characters) considering some homogeneity criterions:
  - Tokens must have only alpha or numerical characters. If there is a change from alpha to number or vice versa, the token must be subdivided.
  - Punctuations characters must be considered as independent tokens

Additionally to these language independent rules, it is possible to add new rules focused on one language or on a set of languages (Romances ones, for example). For example, in English, it is possible to consider subdividing a token if there is a change between lower and upper letters or vice versa (Sproat et al., 2001).

Secondly, some of the tokens (NSWs) are re-written in a different format in order to facilitate their posterior translation. At this step, before rewriting, it is necessary to classify each token as a standard word (W) or as a non-standard word (NSW). This classification can be done considering a dictionary of standard words in this language or considering a more complex classifier based on some features obtained from the target token and its context: character language model, vowels, capitals, etc. In this work, the machine translation module has to deal with this ambiguity without adding additional information.

If the token is classified as a NSW, it is split into letters including some separators at the beginning and at the end of the letter sequence. For example, UPM (Universidad Politécnica de Madrid in Spanish) is rewritten into # U P M #. This way of rewritten an alpha token tries to introduce a high flexibility to facilitate the text normalization process. Considering sequence of letters, some non seen abbreviations could be normalized using the translations of its letters individually.

Also, all the numbers are rewritten dividing the token into digits. Every digit is complemented with its position in the number sequence. For example: 2012 is rewritten as $2_4\ 0_3\ 1_2\ 2_1$, where 2_4 means the digit 2 in the 4th position (beginning from the right). The Roman numbers are first translated into Arabic ones before rewriting.

As it will be shown in the next section, the translation module can deal with graphs of tokens as input. Thanks to this characteristic, it is possible to work with fuzzy decisions when classifying every token as standard word or NSW. Considering a token

graph, both alternatives can be considered with different weight if necessary. **Fig. 2** shows an example of token graph for the sentence "Welcome to UPM2012".

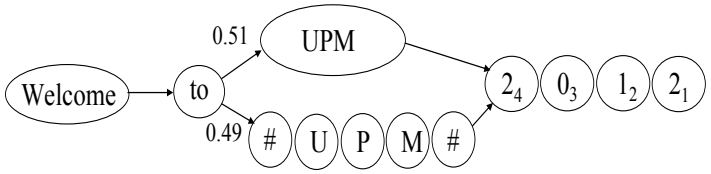


Fig. 2. Token graph for the sentence "Welcome to UPM2012"

The token "UPM2012" is divided into two tokens: UPM and 2012. The first one, UPM, is rewritten considering two possibilities (with two probabilities): as it is, and letter by letter. The second one is a number and it is rewritten digit by digit, including information about its position.

The main target of the standard vs. non–standard word classifier is to detect with high accuracy standard words in order to reduce the token graph complexity, avoiding alternative paths in these cases.

3.2 Token translation

The token translation is performed using a phrase-based system. The phrase-based translation system is based on the software released from Workshops on Statistical Machine Translation (http://www.statmt.org). The Moses decoder is used for the translation process (Koehn et al., 2007). The translation process uses a phrase-based translation model and a target language model.

These models have been trained according to these steps (**Fig. 3**).

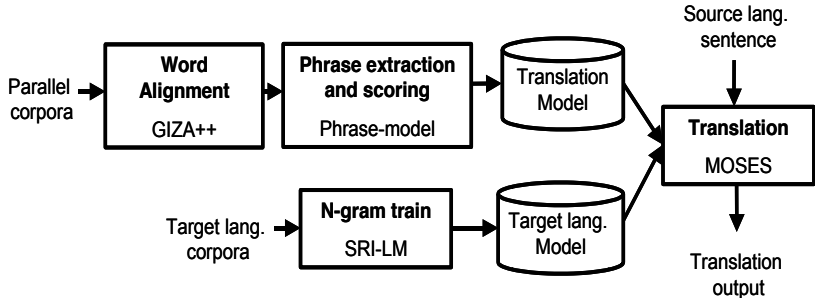


Fig. 3. Process for training the translation and target language models.

- Translation model

In order to generate the token translation model, it is necessary to train a translation and a target language model.

For training the translation model, it is necessary to develop a parallel corpus including examples of all possible typos of NSW described in the taxonomy presented at (Sproat et al., 2001). Some examples are:

- Abbreviations and Acronyms: "The UPM is …" and "The Universidad Politécnica de Madrid is ...".
- Numbers: "more than 120 cars" "more than one hundred and twenty cards".
- Dates and times: "On May 3rd, 2012" "on may third , two thousand and twelve"
- Webs and emails: "example@upm.es" example at U P M dot E S"
- Money and percentages: "$3.4 billions" "three dot four billions dollars"
- Misspelling or funny spelling: "CU8er" "see you later".

This is the most important aspect when developing the text normalization module. The system performance depends strongly on the data used to train the translation model. Also, parallel corpus generation is a costly task that should be supported with automatic procedures to reduce this cost. With this idea, many efforts have been devoted to obtain appropriate corpora from raw text with a small supervision (Collins et al., 1999; Larkey et al., 2000; Sproat, 2010).

One important aspect to consider is that the source language (in the parallel corpora) must be pre-processed in the same way as the input text with the difference that in this case, the parallel corpus does not have token graphs with two alternatives but only token sequences with the correct alternative.

In order to train the translation model, the first step is a word alignment computation. In this step, the GIZA++ software (Och and Ney, 2003) has been used to calculate the alignments between source and target tokens. In order to establish these alignments, GIZA++ combines the alignments in both directions. As there are many standard words, they are the same tokens in source and target languages, being important reference points for the alignment.

The second step is phrase extraction (Koehn et al., 2003). All token phrase pairs that are consistent with the token alignment are collected. For a phrase alignment to be consistent with the word alignment, all alignment points for rows and columns that are touched by a rectangle have to be in the rectangle, not outside. The maximum size of a phrase has been increased to 20 in order to deal with token graphs including sequences of letter and digits properly.

- Target language model

In order to obtain an N-gram language model needed by Moses, the SRI language modelling toolkit has been used (Stolcke, 2002). It is important to consider the target side of the parallel corpora, but also, normalized sentences in different contexts. These additional sentences are interesting to learn the best normalization for a given NSW, depending on the context.

- Translation process

The Moses decoder (http://www.statmt.org/moses/) is used for the translation process. This program is a beam search decoder for phrase-based statistical machine translation models.

One interesting characteristic of Moses is the possibility of combining different phrase tables (translation models) during the translation process using different weights. Considering this possibility, an interesting analysis for future work will be to compare the possibility of training individual phrase tables for each type of NSW or generating a unique one.

3.3 Post-processing

This module performs several actions in order to generate the normalized text to the speech synthesizer. One of the main actions has been to remove unnecessary tokens. For example, if after the translation module there are any # tokens (used for defining the limits of the letter sequences), they must be removed.

Additionally, given that the translation module can generate a token graph or a N-best token sequence, it would be possible to add new translation modules in order to improve the translation process by considering new language models for reordering the N-best token sequences or searching the output token graph.

4 Initial Experiments

In this paper, initial experiments are reported focused on numbers and abbreviations. About numbers, the main target is to define how the architecture can be adapted to deal with numerical numbers in general. The second objective is to deal with abbreviations (including acronyms), distinguishing when a token is a NSW (acronyms or abbreviations) or a standard word. For these experiments, a parallel corpus has been created considering an already developed text normalization module based on rules and word lists. The main idea is trying to learn these rules from data automatically.

For evaluating the performance of the translation system, the BLEU (BiLingual Evaluation Understudy) metric has been computed using the NIST tool (mteval.pl) and the WER (Word Error Rate). It is important to note that BLEU is an accuracy metric while WER is an error metric.

4.1 Experiments with numbers

For these experiments, three parallel corpora have been considered, 800 numbers for training, 1000 for validation and 4000 for testing. These data sets have been created randomly; guarantying that one number only appears in one of the sets. **Table 1** includes some examples from the parallel corpora.

Original text	Normalized test
123.456,34	Ciento veintitrés mil cuatrocientos cincuenta y seis con treinta y cuatro
1.256,3	Mil doscientos cincuenta y seis con tres

Table 1. Examples *of numbers*.

Table 2 shows different experiments considering different codification strategies. In the first one, the digits are grouped in groups of three digits. In this case, there are many errors coming from the confusion between dots referring to millions or thousands. When considering the integer part completely, the results improve significantly. Finally, in the last experiment, a different codification strategy is considered for the decimal part. In this case, the position is coded from the right of the decimal part instead from the comma character: right to left instead of left to right.

System or experiment	BLEU (%)	WER (%)
Baseline: considering groups of three digits Example: 123.400,2 - 1_3 2_2 3_1 . 4_3 0_2 0_1 , 2_1 3_2	80.5	10.3
No considering groups of three digits Example: 123.400,2 - 1_6 2_5 3_4 4_3 0_2 0_1 , 2_1 3_2	96.1	2.2
No considering groups of three digits and different codification for decimals. Example: 123.400,2 - 1_6 2_5 3_4 4_3 0_2 0_1 , 2_{-2} 3_{-1}	**96.6**	**1.9**

Table 2. Different codification strategy for numbers normalization

In order to analyze the effect of the size of the training set, authors performed two additional experiments increasing and reducing the amount of data to train the translation model (**Table 3**):

Different amount of training data	BLEU (%)	WER (%)
400 numbers	92.6	4.4
800 numbers	96.6	1.9
1800 numbers	**97.5**	**1.5**

Table 3. Experiments with different training sets

As it is shown, a good compromise for the training set is around 1000, in order to get a WER lower than 2%.

4.2 Experiments with abbreviations

For these experiments a parallel corpus with 5225 sentences has been divided in training (4054 sentence), tuning (500 sentences), and testing (671 sentences). Every sentence contains one abbreviation (or acronym). **Table 4** includes some examples from the parallel corpora.

Original text	Normalized test
El BBVA subió los precios (The BBVA bank increased the prices)	El be be uve a subió los precios
UGT no negociará más (UGT will not negotiate more)	U ge te no negociará más

Table 4. Examples of sentences with abbreviations.

Table 5 shows the results for the experiments with abbreviations.

Abbreviations Experiments	BLEU (%)	WER (%)
Baseline	96.1	2.9

Table 5. Experiments with abbreviations

The main errors come from those examples that appear in the test set but not in the training set. In these cases, the system leaves the abbreviations as they are generating errors.

5 Conclusions

In this paper, authors have presented initial efforts for developing a text normalization module to be included in a text to speech conversion system. During the design of this module the main characteristics considered have been language independence, based on data instead of expert rules and a high level of flexibility to deal with all situations presented in this task. The architecture proposed in this paper is based on a phrase-based translation system (Moses), considering its main possibilities: dealing with word-graphs at the input and combination of different translation models. This architecture is composed by three modules: a tokenizer module for splitting the text input into a token graph (tokenization), a phrase-based translation module (token translation) and a post-processing module for removing some tokens. Initial experiments with numbers and abbreviations have reported very good results validating the architecture proposed in this paper.

6 Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement n° 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02) and MA2VICMR (CAM, S2009/TIC-1542) projects. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

7 References

1. Sproat, R., Black, A., Chen, S., Kumar, S., Ostendorf, M., and Richards, C., 2001. Normalization of non-standard words. Computer Speech and Language, 15(3), 287-333, 2001.
2. Yeates. S., 1999. Automatic Extraction of Acronyms from Text. New Zealand Computer Science Research Students' Conference. 1999.
3. Larkey, Leah, Paul Ogilvie, Andrew Price and Brenden Tamilio, 2000. Acrophile: An Automated Acronym Extractor and Server, In Proceedings of the ACM Digital Libraries conference, pp. 205-214, 2000.
4. Chang, J.T., Schütze, H., and Altman. R.B., 2002. Cre-ating an Online Dictionary of Abbreviations from MEDLINE. JAMIA.
5. Cannon, G., 1989. Abbreviations and acronyms in English word-formation. American Speech, 64:99-127.
6. Pennell D., and Liu Y., 2011a. Toward text message normalization: Modeling abbreviation generation. Proceedings of the IEEE. pp. 5364-5367.
7. Sproat, R., 2010. Lightly Supervised Learning of Text Normalization: Russian Number Names, IEEE Workshop on Spoken Language Technology, Berkeley, CA, 2010.
8. Bikel, D., Schwartz, R., and Weischedel, R., 1999. An algorithm that learns what's in a name. Machine Learning, 34(1/3):221--231, 1999.
9. Collins, M., and Singer, Y., 1999. Unsupervised Models for Named Entity Classification. EMNLP/VLC-99.

10. Jonnalagadda and Topham. 2010. NEMO: Extraction and normalization of organization names from Pub-Med affiliations. J Biomed Discov Collab. Vol 5 pp 50-75.
11. Brody, S., Diakopoulos. N., 2011. Cooooolllllll!!!!!!!! Using Word Lengthening to Detect Sentiment in Mi-croblogs. EMNLP 2011.
12. Han B., and Baldwin, T., Lexical normalisation of short text messages: Makn sens a #twitter. ACL 2011.
13. Kaufmann, M., 2010. Syntactic Normalization of Twitter Messages. Int'l Conference on NLP.
14. Aw, et al. 2006. A phrase-based statistical model for SMS text normalization. ACL 2006.
15. Pennell D., and Liu. Y., 2011b. A Character-Level Ma-chine Translation Approach for Normalization of SMS Abbreviations. IJCNLP.
16. Och J., Ney. H., 2003. "A systematic comparison of various alignment models". Computational Linguistics, Vol. 29, No. 1 pp. 19-51, 2003.
17. Koehn P., F.J. Och D. Marcu. 2003. "Statistical Phrase-based translation". Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127-133, May 2003.
18. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.
19. Stolcke A. 2002 "SRILM – An Extensible Language Modeling Toolkit". ICSLP. 2002. Denver CO, USA.

GenProso: a parametric prosody prediction module for text-to-speech applications

Juan María Garrido Almiñana

Computational Linguistics Group (GLiCom), Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona
juanmaria.garrido@upf.edu

Abstract. This paper describes GenProso, a module for the prediction of prosodic parameters in text-to-speech (TTS) applications. The paper describes the general architecture of the system, its main modules, the obtained output and its current application.

Keywords. Text-to-speech, prosody prediction, Spanish, Catalan.

1 Introduction

In this paper we present GenProso, a module for the automatic prediction of prosodic parameters (duration and F0) in text-to-speech (TTS) applications developed at the Computational Linguistics Group (GLiCom) of Pompeu Fabra University. It is currently used with TexAFon, the text processing tool of GLiCom, to get the necessary linguistic information from the input text. It has been conceived as a research tool to explore the use of parametric prosodic models for the prediction of prosody in TTS.

Written in Python [1], linguistic/phonetic knowledge in GenProso has been implemented in the form of:

- Python procedures, responsible of the prediction process;
- F0 and duration models, stored in tabular format as text files, which are usually obtained from the analysis of annotated corpora, but can be manually edited by the user for specific research purposes; they can be representative of the speech of one or several speakers, or a particular speech type, depending on the reference corpus used to build them.

The general architecture of GenProso is presented in Figure 1. GenProso receives from TexAFon the phonetic transcription and segmentation in prosodic units (syllables, stress groups and intonation groups) of the input text. The user has also to provide as arguments the particular F0 and duration models that will be used for prediction. These models are built automatically from annotated corpora using two model generation tools also developed at GLiCom.

Next sections present the structure and workflow of these components in more detail.

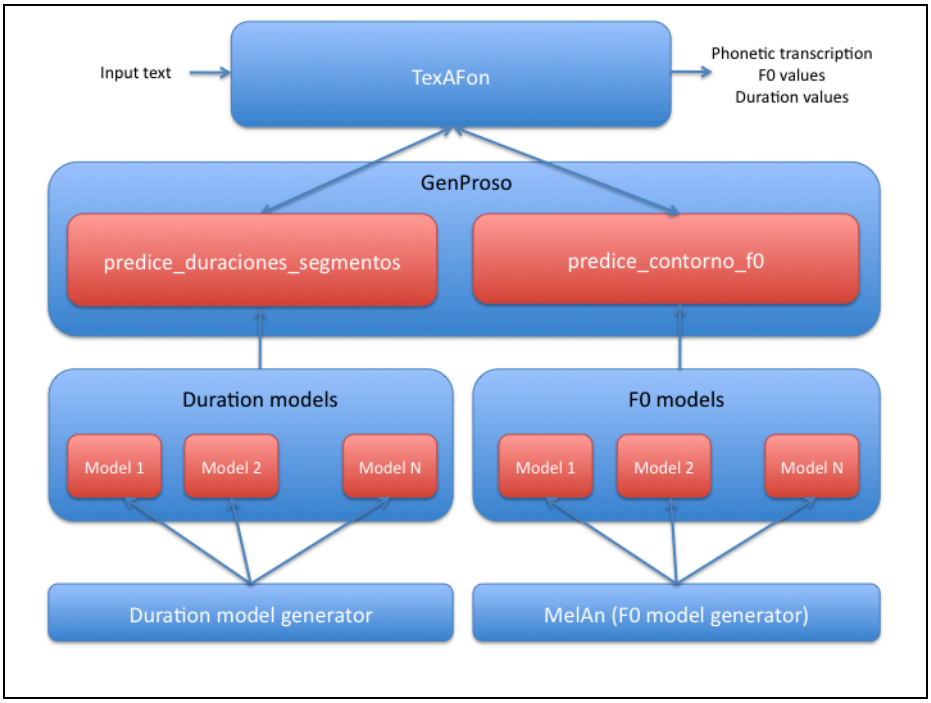


Fig. 1. General GenProso architecture

2 Duration predictor

2.1 Duration models

Duration models are stored in 'txt' files, and include target duration values for the different combinations of considered factors, which are described in table 1. The selected factors have been described in the literature as being relevant to predict the duration of sounds in several languages ([2], [3], [4] among others) and have been widely used in other duration models for TTS previously described ([5], [6], [7], for example). Target values are obtained from the statistical analysis of annotated speech corpora: they are calculated by computing the mean of all the occurrences of the phones sharing the same variable labels in the reference corpus. They are built by running a set of Praat [8] and R [9] scripts which extract duration values from TextGrid files and calculate the corresponding mean values. Table 2 shows an example of one of these models.

Table 1. Variables considered for duration modelling

Label	**Description**
Sonido	Phonetic symbol (including pauses)
Acento	Phone belongs to a stressed or unstressed syllable
Fin_frase	Pause is (or it is not) at the end of a sentence (only applied to pauses)
Silaba_prepausal	Container syllable of phone is (or it is not) before a pause
Modalidad	Mood of the container sentence
Acto_habla	Speech act of the container utterance
Emocion	Emotion label of the container utterance
Nivel_emocion	Emotion level of the container utterance

Table 2. Example of duration model

Modalidad	Acto_habla	Emocion	Nivel_emocion	Acento	Silaba_prepausal	Fin_frase	Sonido	Num_casos	Duracion_media
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	INDEF	INDEF	FIN_FRASE	...	78	0.263
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	INDEF	INDEF	NO_FIN_FRASE	...	148	0.143
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	INDEF	INDEF	PAU_FIN_FRASE	...	213	0.061
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	NO_ACENTUADA	INDEF	INDEF	a	22	0.12
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	NO_ACENTUADA	NO_PREPAUSAL	INDEF	a	1255	0.058
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	NO_ACENTUADA	PREPAUSAL	INDEF	a	121	0.117
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	ACENTUADA	INDEF	INDEF	a_&quot	4	0.080
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	ACENTUADA	NO_PREPAUSAL	INDEF	a_&quot	505	0.091
ENUNCIATIVA	INDEF	ALEGRIA	INDEF	ACENTUADA	PREPAUSAL	INDEF	a_&quot	20	0.097

2.2 Prediction process

The input of the process is a phonetic transcription, enriched with the necessary linguistic information for each phone (stress, prepausal/non prepausal, emotion, etc.). TexAFon predicts this information from the input text, and sends it to GenProso. The prediction process is quite straightforward: for each annotated phone at the input, GenProso looks in the selected duration table the corresponding target value. If no value is found, a general backup value, taken also from the table, is chosen (the one corresponding, for the target phone, to the combination of labels 'NO_ACENTUADA', 'NO_PREPAUSAL' and 'ENUNCIATIVA', which is expected to be always present in the tables).

3 F0 predictor

F0 prediction process assumes the F0 modelling framework described in [10], [11] and [12]. In this framework, F0 contours are considered to be the result of the combination of two different kinds of patterns:

- **Global**, representing the evolution of the contours at Intonation Group (IG) level; in the current implementation of the model, an IG is equivalent to a breath group (a segment of speech between two silent pauses).
- **Local**, predicting the evolution of F0 at Stress Group (SG) level; SG are formed by one stressed syllable and all the following unstressed syllables before the next stressed one (or the end of the IG).

F0 contours are modeled as series of relevant inflection points, representing relevant changes in the direction of the F0 along the contour. Each inflection point is assigned to the P (Peak) or V (Valley) level, considering its relative height within the F0 range of its container IG. Figure 2 shows an example of this kind of representation. Two extra labels, P+ and V-, are used to mark inflection points showing F0 values clearly higher or lower than the P and V mean levels, respectively.

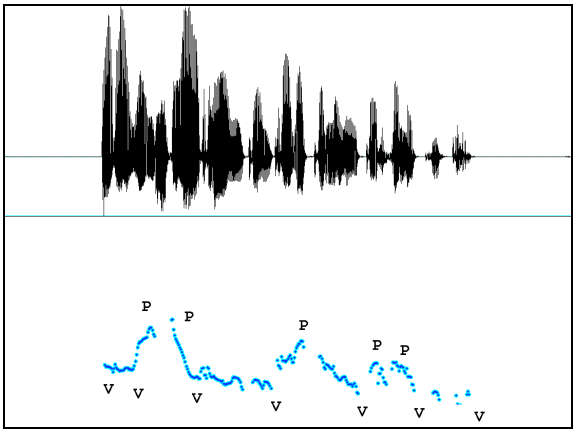


Fig. 2. Waveform and F0 contour of the utterance "*Aragón se ha reencontrado con el motor del equipo*", uttered by a Spanish female speaker.

Global patterns are represented in the model as 'reference lines' showing the evolution of the P and V F0 levels along the IG. For each IG, then, two reference lines are considered, one for the P and one for the V level. These patterns are speaker-dependent, and model the F0 height and range of each speaker. Figure 3 illustrates this concept.

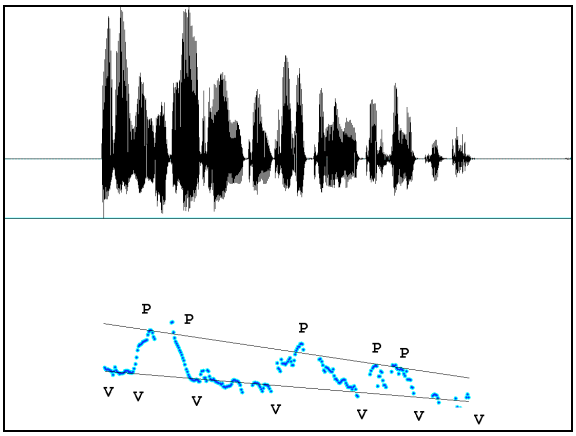


Fig. 3. Waveform and F0 contour of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker.

Local patterns are defined as recurrent series of labels anchored at specific places of the syllables that make up SG. The position of each point is defined with respect to the nucleus of its container syllable. Three different positions are considered: I ('initial', close to the beginning of the syllable nucleus), M ('middle', close to the centre of the nucleus), and F ('final', close to the end of the nucleus). Figure 4 shows an example of such kind of patterns. Vertical lines represent the boundaries between SG, delimiting the different local F0 patterns of the contour.

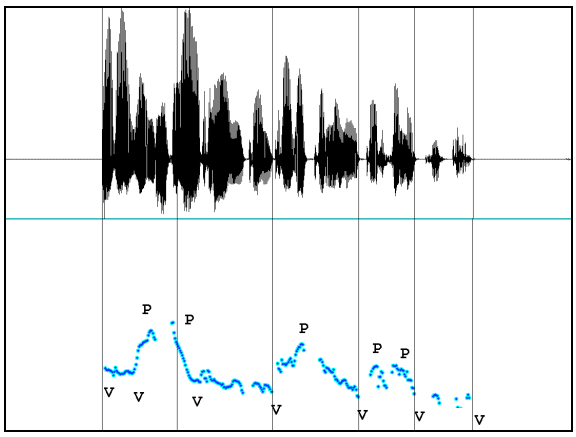


Fig. 4. Waveform and F0 contour of the utterance "*Aragón se ha reencontrado como motor del equipo*", uttered by a Spanish female speaker. Vertical lines represent SG boundaries.

3.1 F0 models

According to this framework, two kinds of models are used in GenProso for the prediction of F0 contours: those predicting global patterns, modelled as prototypical

reference lines for the different types of considered IG; and those for the local patterns, represented as series of inflection points anchored to specific places of the container SG.

Global pattern models are tables, stored in 'txt' files, in which each row represents a global pattern for a type of IG. Each pattern includes the initial and final F0 values for P and V reference lines. Each IG type is defined as a function of the variables described in table 3. An example of such global models is provided in table 4.

Local pattern models are also tables contained in 'txt' files in which each line represents one pattern for a given SG type. SG types are defined by the combination of variables presented in table 5. In this case, several patterns (up to 5) can be provided for the same SG type.

As in the case of duration, both types of F0 models are derived from the statistical analysis of annotated speech corpora: P and V lines for the different IG types are obtained by calculating the mean of the initial and final F0 values of the reference lines of all the particular IG of a given type in the reference corpus; and SG patterns are the result of calculating the frequency of each pattern in the corpus, and retaining the most frequent patterns (up to 5) for each SG type. Current implementation of GenProso only takes into account the most frequent pattern, but future improvements are expected to use this information to offer more intonational variation during synthesis. Table 6 offers an example of the appearance of such local pattern models.

F0 models are built by running MelAn, a fully automatic modelling tool developed at GLiCom [12]. It is made up of a set of Praat and R scripts that perform the following tasks:

- **Stylisation**: only relevant inflection points of the F0 contours are retained:
- **Annotation**: inflection points are assigned to one of the four considered F0 levels (P+, P, V, V-);
- **Modelling**: models are built form the analysis of the global and local patterns in the reference annotated corpus.

Table 3. Variables considered for global F0 patterns modelling

Label	Description
Posicion_GF	Position of the IG within the sentence
Modalidad	Mood of the container sentence
Acto_habla	Speech act of the container utterance
Emocion	Emotion label of the container utterance
Nivel_emocion	Emotion level of the container utterance

Table 4. Example of table for global F0 patterns

Modali-dad	Acto_habla	Emocion	Nivel_emocion	Posi-cion_GF	Num_casos	Valor_ini-cial_P_media	Valor_final_P_media	Pendi-ente_P_media	Valor_ini-cial_V_media	Valor_fi-nal_V_media	Pendi-ente_V_media
ENUNCIATIVA	INDEF	INDEF	INDEF	FINAL	404	141.82	91.15	-37.55	109.07	69.21	-29.32
ENUNCIATIVA	INDEF	INDEF	INDEF	INICIAL	1946	149.52	109.80	-22.92	112.97	85.06	-12.05
ENUNCIATIVA	INDEF	INDEF	INDEF	INICIAL_FINAL	363	158.91	90.57	-45.30	120.27	69.18	-33.74
ENUNCIATIVA	INDEF	INDEF	INDEF	INTERIOR	1823	137.50	106.15	-18.17	108.13	80.72	-14.78
EXCLAMATIVA	INDEF	INDEF	INDEF	FINAL	2	238.59	62.06	-157.77	129.81	84.67	-44.31
EXCLAMATIVA	INDEF	INDEF	INDEF	INICIAL	84	181.17	124.05	-49.21	134.53	86.06	-43.55
EXCLAMATIVA	INDEF	INDEF	INDEF	INICIAL_FINAL	52	204.94	118.99	-93.99	141.13	86.45	-49.36
EXCLAMATIVA	INDEF	INDEF	INDEF	INTERIOR	5	186.84	99.35	-81.39	139.53	80.26	-55.740

Table 5. Variables considered to model local F0 patterns

Label	**Description**
Posicion	Position of the SG within the container IG and sentence
Num_silabas	Number of syllables of the SG
Num_sil_acentuada	Position of the stressed syllable within the SG
Modalidad	Mood of the container sentence
Acto_habla	Speech act of the container utterance
Emocion	Emotion label of the container utterance
Nivel_emocion	Emotion level of the container utterance

Table 6. Example of table for local F0 patterns

Patron	Num_silabas	Num_sil_acentuada	Modalidad	Emocion	Acto_habla	Posicion	Freq
VF0	1	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	55
VI0_VF0	1	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	36
PI0_VM0	1	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	35
VM0	1	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	30
VI0_VM0	1	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	18
VM1	2	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	32
VI1_VM1	2	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	26
PI0_VI1_VM1	2	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	23
VI0_VM1	2	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	20
PI0_VF0_VM1	2	1	ENUNCIATIVA	INDEF	INDEF	FINAL_ENUNCIADO	19

3.2 Prediction process

The building of F0 contours is quite similar to the duration prediction process. In this case, the linguistic information necessary as input is a chain of IG, each one segmented into SG and properly labelled. This information is provided by TexAFon after

analysis of the input text. With this information, GenProso calculates the target F0 contour for each input IG. This calculation process involves two steps:

- **Pattern selection**. First the global F0 pattern fitting the input IG type is selected from the table. If no pattern is found for the required IG type, a general backup pattern is chosen (the one corresponding to the combinations of labels 'FINAL' and 'ENUNCIATIVA', which is expected to be always present in the tables). Then, for each SG in the IG, the corresponding local F0 pattern is selected. If no pattern is found for the required SG type, a general backup pattern is chosen (the one corresponding to a one-syllable SG in the target SG position, which is expected to be always present in the tables).
- **F0 turning points calculation**. GenProso converts to particular F0 values the labels (P, V, P+, V-) of the local F0 patterns, by using the selected global reference lines. If the turning point is P or P+, P reference line is used; if not, V reference line is selected. Particular F0 values are obtained by calculating the F0 value at the corresponding time value of the turning point in the IG. If the point is labelled as P+, the obtained F0 value is raised by a fixed percentage; if it is V-, it is lowered also accordingly.

4 Output

The output of GenProso is twofold: a chain of pairs [phonetic symbol, duration]; and a chain of pairs [phonetic symbol, F0 values]. Up to three F0 values can be associated to each input phone, at initial, mid or final position. This output is returned to TexAFon, which uses it to generate the corresponding 'pho' file to be sent to the Mbrola engine [13], which is currently used to obtain the output synthetic speech.

5 Evaluation

A perception test was carried out to evaluate the quality of the prosody predicted by GenProso. 'Pho' files for 10 selected utterances in Spanish were generated using TexAFon + GenProso. Table 7 presents the text of these utterances.

'Wav' files were obtained from these files by means of Mbrola. These files were used to prepare the perception test, accessible online. 10 subjects, all of them mother tongue speakers of Peninsular Spanish, were asked to rate the quality and naturalness of the prosody of each utterance in a 1 to 5 scale (5 being the best possible qualification). They were encouraged to judge exclusively the intonation of the stimuli, not their overall quality (obviously limited due to the use of Mbrola). They were also encouraged to run the test in a quiet room, and to use headphones, but the running was not supervised by the experimenter. Table 8 presents the obtained results: mean score value was 3.44, which suggest a moderate acceptance of the generated prosody, in terms of intelligibility and naturalness, by the subjects.

Table 7. Utterances used in the evaluation test

Declarative	*El servicio de información deportiva de Radio Nacional tiene un precio de cero coma treinta céntimos por llamada.*
	Abría sus puertas a estos flacos alumnos afroamericanos.
	Valencia, Murcia y Andalucía siguen en alerta ante las intensas lluvias.
	En la operación, enmarcada en la lucha contra el crimen organizado, se han efectuado registros en once viviendas de Blanes y Girona.
Interrogative (total)	*¿Tomará usted una copita?*
	¿Desea añadir el nombre del contacto a su libreta de direcciones?
Interrogative (partial)	*¿Dónde estabas tú en aquel verano del cincuenta y seis?*
	¿Qué piensas hacer allí?
Exclamative	*¡Me has hecho daño!*
	¡Fuera de aquí inmediatamente!

Table 8. Mean scores obtained in the evaluation test

	Utterance										
Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**
1	4	2	4	4	4	2	5	4	5	3	3.7
2	3	5	5	5	4	4	2	3	4	2	3.7
3	3	2	4	3	3	4	3	2	2	3	2.9
4	4	3	4	3	5	2	2	4	2	3	3.2
5	3	4	4	2	3	4	4	4	3	3	3.4
6	4	2	5	4	5	4	3	3	5	4	3.9
7	2	5	3	4	4	4	2	3	4	4	3.5
8	3	1	2	5	4	2	1	2	5	4	2.9
9	4	4	4	3	3	4	4	3	5	4	3.8
10	2	3	3	2	4	3	5	2	5	5	3.4
Mean	3.2	3.1	3.8	3.5	3.9	3.3	3.1	3	4	3.5	**3.44**

6 Concluding remarks

The obtained evaluation results seem to indicate that GenProso is able to generate prosody with an acceptable level of naturalness and quality. It seems to be an adequate research tool to explore the use of parametric models for the generation of synthetic speech in different situations (emotional, dialogues, etc), as it is expected to be done in the near future.

7 References

1. Python Language Website, http://www.python.org/.
2. Navarro, T.: Manual de pronunciación española. Madrid: Consejo Superior de Investigaciones Científicas (1918)
3. Klatt, D.H.: Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, 59, 5, pp. 1208-21 (1976)

4. Borzone de Manrique, A.M., Signorini, A.: Segmental duration and rythm in Spanish. Journal of Phonetics, 11, pp. 117-128 (1983)
5. Klatt, D.H.: Synthesis by rule of segmental durations in English. In Lindblom, B., Öhman, S. (eds.): Frontiers of Speech Communication Research, London: Academic Press, pp. 287-299 (1979)
6. van Santen, J.: Deriving text-to-speech durations from natural speech. In Bailly, G., Benoit, C. (eds.): Proceedings of the ESCA Workshop on Speech Synthesis, pp. 157-160 (1990)
7. Macarrón, A., Escalada, G., Rodríguez, M. A.: Generation of duration rules for a Spanish text-to-speech synthesizer. In Proceedings of Eurospeech'91, pp. 617-620 (1991)
8. Boersma, P. and Weenink, D. Praat: doing phonetics by computer [Computer program] http://www.praat.org/ (2012)
9. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org. (2008)
10. Garrido, J. M.: Modelling Spanish Intonation for Text-to-Speech Applications. Ph.D. thesis, Universitat Autònoma de Barcelona (1996)
11. Garrido, J. M.: La estructura de las curvas melódicas del español: propuesta de modelización. Lingüística Española Actual, XXIII/2, pp. 173-209 (2001)
12. Garrido, J. M.: A Tool for Automatic F0 Stylisation, Annotation and Modelling of Large Corpora. In: Speech Prosody 2010, Chicago, May 2010. (2010)
13. Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vrecken, O.: The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: Proc. ICSLP 96, Philadelphia, 3, pp. 1393-1396 (1996)

MAVIR: a corpus of spontaneous formal speech in Spanish and English

Antonio Moreno Sandoval and Leonardo Campillos Llanos

Laboratorio de Lingüística Informática (LLI), Universidad Autónoma de Madrid,
28049 Madrid, SPAIN
{antonio.msandoval, leonardo.campillos}@uam.es

Abstract. MAVIR corpus is a collection of audio and video recordings, with their corresponding orthographic transcriptions and prosodic annotation. The main aim of the corpus is researching in Natural Language Processing and Speech Technology. Recordings come from lectures and talks on language technologies celebrated within the framework of MAVIR consortium. The corpus is made up of 13 recordings (audio and video) in Spanish and English languages, collected during the I, II and III MAVIR Conferences, held in Madrid in 2006, 2007 and 2008 respectively.

Keywords: Language resources, spontaneous speech, formal speech.

1 Introduction

Spoken language resources are indispensable data for developing and evaluating speech systems. In this paper, we describe a collection of audio and video recordings of formal, spontaneous, speeches in Spanish and English. Those recordings were taken in a series of lectures and panels organized by the MAVIR consortium[1] between 2006 and 2008. Spoken language resources are typically divided into speech databases and spontaneous speech corpora. The former are collections of high-quality recordings and detailed phonetic transcriptions of speech in controlled environments. The later are typically collections of a wide variety of spoken registers and non-scripted speech. Those corpora are collected mainly for linguistic analyses and applications such as language teaching, or writing grammars and dictionaries. The first spoken corpora collected were part of general, reference, national corpora such as BNC [2] or CREA [4].

Spanish research groups have not been very active in the compilation of spontaneous speech processing. Analogously, there are a few corpora available for spontaneous speech in Spanish [1,5]. The LLI-UAM has a long history of such as resources.

[1] MAVIR: Mejorando el Acceso y Visibilidad de la Información en Red (http://www.mavir.net/) is a research consortium funded by the Madrid Regional Government under the grants S0505/TIC-0267 and S2009/TIC-1542.

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Starting in 1990-92, they collected the first spontaneous speech corpus of Spanish, CORLEC [8][2]. A decade later, with a new team, they were responsible for developing the Spanish corpus within the European project C-ORAL-ROM [5][3]. This project, along with its contemporary Dutch Spoken Corpus [6], was conducted during the early years of the past decade. The two projects were an improvement over the aforementioned national corpora in various aspects detailed in [11]:

1. The acoustic quality: from analog tapes to digital recording. In the CORLEC times, simply there were not digital recorders. One of the aims of the C-ORAL-ROM project was to provide data to the language technology community with sufficient quality.
2. Clear separation of the metadata (header), from the text transcription.
3. The synchronization of transcription and audio (by utterances). This is useful to segment the signal according to the text, but also to check the quality of transcription.
4. Prosodic and Part-of-Speech annotations were provided in different layers.
5. Legal rights of the speakers and copyrights holder (in media recordings) are preserved. Every recording has the written permit from the participant.

Therefore, a clear evolution of the oral corpus can be seen, on the basis of maintaining the essence: recording of spontaneous speech in its context of use.

2 Description of the MAVIR Corpus

MAVIR corpus has been constructed under the experience of previous corpora, C-ORAL-ROM and CHIEDE[4] [7] but MAVIR is a bilingual corpus (Spanish and English) with important differences with respect to those mentioned (see Table 1).

Table 1**.** Distinctive features of the three corpora.

	C-ORAL-ROM	**CHIEDE**	**MAVIR**
General type	Reference corpus	Child corpus	Topic-oriented
Design	Formal vs. informal	By child ages	By topic
Interactional type	Monologues, dialogues, conversations	Dialogues and conversations	Monologues and panel discussions
Typical length of recordings	Between 5 and 30 minutes	Between 10 and 30 minutes	Between 20 minutes and one hour

[2] http://www.lllf.uam.es/ING/Corlec.html. The transcription is available for downloading.
[3] http://www.lllf.uam.es/ING/Coralrom.html
[4] http://www.lllf.uam.es/ING/Chiede.html

MAVIR is a corpus of formal speech, in contrast with C-ORAL-ROM and CHIEDE, where the informal speech is the characteristic feature. Another relevant aspect is its topic orientation: lectures on language technologies issues such as information retrieval or semantic web. The Table 2 shows the text distribution by language and topic.

In figures, MAVIR consists of 13 files (9 in Spanish and 4 in English) with a total duration of more than 10 hours and over 100,000 words, including 3 hours and 10 minutes in English and over seven hours in Spanish (table 2).

Table 2. MAVIR text distribution

File	Title	Duration	Nº of words[5]	Nº of utterances	Lang.
mavir01	Challenges for Information Extraction	1h 07' 39"	9113	597	Eng
mavir02	Proceso de innovación de tecnologías de acceso a la información: ¿Cómo llegar al mercado?	1h 14' 32"	13422	682	Spa
mavir03	España y los buscadores: un mercado potencial	38' 11"	6681	481	Spa
mavir04	Aplicaciones en dominios médico y cultural	57’ 22"	9310	347	Spa
mavir05	On-demand Information Extraction	36' 08"	4461	464	Eng
mavir06	Buscador General Panhispánico	29' 09"	4332	140	Spa
mavir07	Tecnología de la Web Semántica	21' 47"	3831	190	Spa
mavir08	Premio MAVIR 2007	18' 55"	3356	189	Spa
mavir09	Buenas prácticas en presencia web para grupos de investig.	1h 10' 03"	11179	650	Spa
mavir10	Multimedia Retrieval and Evaluation	1h 27' 24"	15659	657	Eng
mavir11	Premio MAVIR 2008	20' 20"	3130	152	Spa
mavir12	Beyond Text-based Multimedia Retrieval	1h 7' 40"	11168	741	Eng
mavir13	Buscando cangrejos en Flickr	43' 38"	7837	531	Spa
TOTAL		10h 38' 48"	103479	7902	

[5] The word count has been provisionally performed considering every item between two spaces; so, actually, a multiword such as *es decir* (‘that is’) counts as two words.

The total number of words for each language is 63078 in Spanish and 40401 in English. With respect to participants, the four texts in English are monologues, while the seven Spanish-language recordings are split between monologues and round tables, with a total of 19 different speakers.

3 Methodology

The recordings were taken at the scene of conferences (in different sites). In most cases the signal was taken directly from the audio system. In other cases, the lectures were recorded with a DAT recorder. Speech signal was down-sampled to 16kHz, 16-bit mono.

For editing recordings, we used the program CoolEdit©. This software allows one to manipulate the sound, thus improving quality, eliminating noise or cutting out parts that are not relevant.

The corpus was transcribed and prosodically annotated by several transcribers, all of them Ph.D. students with a background in linguistics. Experienced members of the LLI-UAM supervised the whole task. The transcribers based their annotation on the transcription guidelines, following the C-ORAL-ROM conventions [10]. Each transcriber made a first version, which was revised by another transcriber; after revision, they discussed disagreements and reached a final version.

Transcriptions were carried out from the processed sound files. The text of a transcription is divided into two parts: header and transcription. Information regarding the participants and the communicative situation is included in the header; for instance, speaker's data (sex, education, dialect, etc.), topic, duration, transcribers or revisers.

The second part of the file, after the metadata section, is the text transcription, carried out through the orthographic transliteration of the recordings, following specific conventions developed specifically for spoken language. The punctuation system established for written language is not suitable for spoken language. Next, we will describe briefly the conventions.

4 Transcription conventions

Figure 1 below shows an example of the transcription of recording mavir05.

Figure 1. Fragment of the transcription from file mavir05

*SEK: ok /// so this is the result /// you can tell /// right ? this is a result xxx was supposed to get /// I can pick one of them /// maybe this one /// {%com: he waits until the page loads} Netherlands beats Spain /// hhh {%act: interjection} beat hhh {%act: laugh} I didn't know /// you know what I'm forward to xxx /// so / yeah ? this is what / maybe / &ah we can expect from / a question like country name ///

Following is a summary of the transcription conventions used in the corpora.

Table 3. Transcription conventions

Mark	Description	Meaning	Example
/	Non-terminal pro-sodic break.	Non-autonomous tonal units	*GRI: thank you Antonio / and thank you (…)
//	Non-terminal auto-nomous pros. break.	Independent tonal units	*GRI: ok // so for exam-ple / &ah (…)
///	Terminal prosodic break	Informative units (complete meaning)	*SEK: that's the idea ///
¡!	Exclamat. utterance	Exclamation	*GRI: this problem got solved !
¿?	Interrogat. utterance	Interrogation	*SEK: you get idea ?
…	Not-finished utterance	Suspended intonation	*SEK: but / at the moment ...
=	Self-interruption	Intentional interruption	*GRI: here = ups! / excuse me
+	Interruption	Speaker is interrupted.	*LRO: el caso de xxx +
¬	Turn continued after an interruption	It is used at the beginning of the interrupted turn.	*IRA: millones / *ENR: no /// no /// *IRA: ¬ se gastaron
→	Lengthening	Long vowel/conson.	*SEK: all → one thousand
[/]	Simple retracting	Repetition or retrace	*GRI: not [/] no job get started
[///]	Retracting	Syntactic reformulation	*GRI: Booth was &assassina [///] sorry ///
< >	Overlapping	It is used when two people speak at the same time.	*JSL: <es capaz de resolvérselo> *LRO: [<] <de resolvérselo>
#	Non-prosodic break	A long break (not expressive intention)	*COR: herramientas sencillas # {%com: consults laptop}
xxx	Not-transcribed words	Passage not understood	*GRI: literature xxx
&	Before a fragment or unfinished word	A non-complete word (self-correction)	*GRI: so being &ab [/] being able to pull out
&eh &ah &mm	Vocalic support or filler	The speaker uses it to keep his / her turn.	*GRI: &eh similar obser-vation can be made
hhh {%act:}	Paralingüistic or non-linguistic elem.	An onomatopoeia, laugh, assent, click…	*GRI: hhh {%act: cough}
{%alt:}	Production errors	A wrong word or mispronunciation.	*SEK: &ah promotion {%alt: promo-tion} to xxx
{%com:}	Comments	It comments an event	*GRI: {%com: drinks}

5 Alignment

The alignment involves the text synchronization with the original sound, either by conversational turns or by utterances (in our case, the corpus is aligned by utterances).

In the first stage, every text fragment is synchronized with the corresponding sound. Trained linguists manually carry out this work by means of professional software, and it is a painstaking task, since it requires precision when marking the initial and the final time codes for every utterance. After the synchronization is finished, the conversion to the XML format is automatically performed. The text is broken down into utterances (according to the time codes marked by the linguist), which are limited by a time stamp at the beginning and at the end of each fragment. Figure 2 shows the synchronized transcription corresponding to the fragment in figure 1.

Figure 2. Fragment from an XML file (mavir05) with the transcription and the time codes

```
<UNIT speaker="SEK" startTime="543.109" endTime="544.578"> ok </UNIT>
<UNIT speaker="SEK" startTime="544.578" endTime="545.99"> so this is the result </UNIT>
<UNIT speaker="SEK" startTime="545.99" endTime="547.125"> you can tell </UNIT>
<UNIT speaker="SEK" startTime="547.125" endTime="548.962"> right ?</UNIT>
<UNIT speaker="SEK" startTime="548.962" endTime="552.248"> this is a result xxx was supposed to get </UNIT>
<UNIT speaker="SEK" startTime="552.248" endTime="554.452"> I can pick one of them </UNIT>
<UNIT speaker="SEK" startTime="554.452" endTime="555.572"> maybe this one </UNIT>
<UNIT speaker="SEK" startTime="555.572" endTime="565.385"> {%com: he waits until the page loads} Netherlands beats Spain </UNIT>
<UNIT speaker="SEK" startTime="565.385" endTime="568.103"> hhh {%act: interjection} beat hhh {%act: laugh} I didn't know </UNIT>
<UNIT speaker="SEK" startTime="568.103" endTime="570.489"> you know what I'm forward to xxx </UNIT>
<UNIT speaker="SEK" startTime="570.489" endTime="571.605"> so / yeah ?</UNIT>
<UNIT speaker="SEK" startTime="571.605" endTime="576.171"> this is what / maybe / &amp;amp;ah we can expect from / a question like country name </UNIT>
```

6 Conclusions

With regard to applications, the MAVIR corpus has been applied to date for the following tasks:

1. A descriptive study of speech dysfluencies of Spanish in formal register [3]

2. Development, training and test of several ASR systems –among them, the AVTS and the THALES-UPM system–. Besides, researchers working in the European project transLectures [13] have been interested in the corpus, and they have been given a copy of the resource
3. The corpus will be put to use in the wordspotting test task which is going to be hold during IberSpeech 2012 conference. The LLI-UAM group has manually annotated more than 5000 words to be used in this competitive evaluation.

The MAVIR corpus is a contribution to resources for the speech technology research community. The corpus is freely available for research purposes. Please contact with Dr. Antonio Moreno Sandoval to get a copy of the DVDs (antonio.msandoval@uam.es).

References

1. Benedí, J.-M., Lleida, E., Varona, A., M.-J. Castro, Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In Proc. of Fifth International Conference on Language Resources and Evaluation, LREC 2006. Genova, Italy.
2. British National Corpus. http://www.natcorp.ox.ac.uk/
3. Campillos, L., Alcántara, M.: Speech Dysfluencies in Formal Context. Analysis based on Spontaneous Speech Corpora. In Proc. Corpus Linguistics Conference 2009 (2009)
4. Corpus de Referencia del Español Actual. http://corpus.rae.es/creanet.html
5. Cresti, E., Moneglia, M., Bacelar do Nascimento, F., Moreno-Sandoval, A., Veronis, J., Martin, P., Choukri, K., Mapelli, V., Falavigna, D., Cid, A., Blum, C.: The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus. In Proc. of Language Resources and Evaluation Conference 2002. Las Palmas, Spain. (2002). C-ORAL-ROM official web site: http://lablita.dit.unifi.it/coralrom
6. Dutch Spoken Corpus. http://lands.let.ru.nl/cgn/ehome.htm
7. Garrote Salazar, M.: CHIEDE. Corpus de habla infantil espontánea del español. Ph.D. Thesis. Madrid: UAM Publishing Service. (2008)
8. González Ledesma, A., De la Madrid, G., Alcántara Plá, M., De la Torre, R., Moreno-Sandoval, A.: Orality and Difficulties in the Transcription of Spoken Corpora. In Proc. of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, 2004, Lisbon (2004)
9. Marcos Marín, F.: El Corpus Oral de Referencia de la Lengua Española contemporánea. Project Report. Madrid (1992). http://www.lllf.uam.es/ESP/Info%20Corlec.html
10. Moneglia, M.: The C-ORAL-ROM resource. In Cresti, E., Moneglia, M. (eds.) p. 27 (2005)
11. Moreno Sandoval, A.: La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. Actas de las Segundas Jornadas de Tecnologías del Habla. Granada (2002)
12. Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, JM., de la Torre, R.: The Spanish Corpus. In Cresti, E., Moneglia, M. (eds.) C-ORAL-ROM: Integrated reference Corpora for Spoken Romance Languages, pp. 135-161. Amsterdam: John Benjamins (2005)
13. TransLectures project (Transcription and Translation of Video Lectures). http://llach.dsic.upv.es/~translectures/

Velar Movement in European Portuguese Nasal Vowels

Paula Martins[1], Catarina Oliveira[1], Samuel Silva[2], and António Teixeira[2]

[1] ESSUA/IEETA, University of Aveiro, 3810–193 Aveiro, Portugal
[2] DETI/IEETA, University of Aveiro, 3810–193 Aveiro, Portugal
pmartins@ua.pt, coliveira@ua.pt, sss@ua.pt, ajst@ua.pt

Abstract. Given the dynamic pattern of the European Portuguese (EP) nasal vowels (i.e. gradual change from an oral configuration towards a nasal one), the acquisition of information on the velum movement over time is of extreme relevance. This study aims to analyse velar dynamics during the production of EP nasal vowels by using real-time resonance imaging (RT-MRI). Audio synchronized coronal oblique slices (at velum port) were acquired from three female speakers, at a frame rate of 14 frames/s. Semi-automatic segmentation techniques were applied, in order to obtain the variation of nasal/oral areas over time, maximum nasal areas and duration of opening-closing velar gestures. The results show similar nasal areas for the different EP vowels, suggesting subtle differences in the magnitude of velum lowering. A gradual increase of nasalization over the vowel was also observed. Duration patterns reveal that the opening movement is longer than the closing gesture, which confirms previous accounts for EP.

1 Introduction

Nasal vowels are produced by a lowering of the velum, so that the nasal cavities are coupled with the oral tract. The movement of the velum during the production of nasal vowels has been studied, in great depth, over the last decades using a varied set of direct (e.g. fiberoscopy [1], cineradiography [2], electromagnetic articulography [3], magnetic resonance imaging [4]) and indirect techniques (e.g. measurements of oral and nasal air flows [5], acoustic analysis [6]). Most of this research has focused on French nasal vowels, but differences between Portuguese nasalization and nasalization in other languages warrant specific studies of Portuguese nasal sounds.

The articulation of European Portuguese (EP) nasal vowels has received a considerable amount of attention by our team, by exploring electromagnetic articulography (EMA) data [7–10] and, more recently, magnetic resonance imaging (MRI) techniques [11, 12].

EMA studies on EP nasal vowels examined essentially duration, velocity (stiffness), magnitude and timing of velum gestures and suggest the following conclusions: **i)** the opening and closing movement of the velum takes about 220ms (fast speech rate) - 340ms (normal speech rate) and the opening movement

is longer than the closing gesture [13, 10]; **ii)** stiffness differences between the opening and the closing gestures of the velum entail differences on the duration of articulatory movements [10]; **iii)** the velum gesture is somewhat delayed with respect to the tongue body gesture, inducing an oral onset, and depending on the degree of overlap between the closing velum gesture and the following oral gesture (as in "c*an*to"), a transitional/intrusive consonantal segment can appear [8, 10]; **iv)** speech rate reduces the duration of velum gestures and increase the stiffness and the intergestural overlap [10]; and **v)** as previous studies on velum height for nasal segments [7, 9] provided some inconclusive or ambiguous results, further data should be acquired to find whether the velum position is identical between the five EP nasal vowels.

Static 2D MRI has been used to obtain information about articulatory differences in the tongue and velum position between oral and nasal vowels, through superimposition of midsagittal contours [11, 12]. 3D MRI allowed us to measure nasal and oral cross-sectional areas and to calculate the velum port opening quotient (VPOQ) [11, 12].

While EMA provides high temporal resolution measurements of articulator movement, although limited to a few points, MRI gives accurate information on the position of the articulators, such as tongue and velum, but with some well-known drawbacks, especially long acquisition times, requiring the subject to artificially sustain the articulation. Thus, both EMA and static MRI have to be complemented with other methods, such as real-time MRI (RT-MRI), in order to fully characterize articulatory movements. Given the dynamic pattern of Portuguese nasal vowels (i.e. gradual change from an oral configuration towards a nasal one), first mentioned by Lacerda and Head [14] and underlined also by perceptual [15] and articulatory studies [8, 10], the acquisition of dynamic information is particularly relevant.

This study complements previous articulatory analysis by using RT-MRI to analyse temporal course of velar movement during the articulation of EP nasal vowels by three EP speakers. The present paper reports, for the first time, variation of nasal and oral areas through time, as well as duration of velar opening and closing movements measured from the nasal areas. Furthermore, potential differences in the magnitude of velum opening between the five EP nasal vowels will be examined, in order to establish if there is more than one position of velic lowering (cf. Demolin et al. [16] for French nasal vowels).

Information concerning areas and durations is crucial to realistically represent the movement of articulatory gestures in our articulatory based text-to-speech system [17].

This paper is organized as follows: section 2 details methods of MR image acquisition and describes the tools used for image processing and analysis; section 3 provides some results; finally, in section 4, we briefly discuss our results and summarize directions for future work.

2 Methods

2.1 Corpus and Speakers

Three female speakers (SV, CM, CO) of EP, aged between 21 and 33, participated in the RT-MRI experiment. None of the speakers reported having hearing or speech disorders. An MRI screening form and informed consent was obtained before the study to comply with security and ethics rules.

The corpus considered in this study represents a subset of a large database (refer to [18] for more details).

The speakers produced nonsense words containing EP nasal ([ɐ̃], [ɛ̃], [ĩ], [õ], [ũ]) vowels, uttered in three prosodic conditions: word-initial, word-internal and word-final (e.g. "ampa, pampa, pan", [ɐ̃pɐ], [pɐ̃pɐ], [pɐ̃]). The nasal vowels were flanked by a voiceless bilabial stop [p].

Audio was recorded simultaneously with the RT-images, inside the MR scanner, at a sampling rate of 16000 Hz, using a fiberoptic microphone.

2.2 Image Acquisition

The MRI experiment was carried out in a Magnetic Resonance Imaging Unit at Coimbra (IBILI – Institute of Biomedical Research in Light and Image). The images were acquired on an unmodified 3.0 T MR scanner (Magneton Tim Trio, Siemens, Erlanger, Germany) equipped with high performance gradients (Gmax = $45mT/m$, rise time = $0.2s$, slew Rate = 200 $T/m/s$, and FOV = 50 cm). A 12-channel head and 4-channel neck phased-array coils were used for data acquisition. Parallel imaging (GRAPPA 2) together with magnetic field gradients operating at FAST mode were used to speed up the acquisition. The subject lay supine in the MR scanner, while producing the stimuli and wore headphones to protect the ears from the noise. The imaging protocol was an evolution of a former study conducted by Martins et al. [19].

After localization images, a T1 W 2D-coronal oblique MRI slice was taken at the velum, using an Ultra-Fast RF-spoiled Gradient Echo (GE) pulse sequence (Single-Shot TurboFLASH), with a slice thickness of 8 mm and the following parameters: TR/TE/FA = $72ms/1.02ms/5^{o}$, Bandwidth = 1395 $Hz/pixel$, FOV(mm^2)= 210 × 210, reconstruction matrix of (128 × 128) elements with 50 % phase resolution, in-plane resolution (mm^2) = 3.3 × 1.6, yielding a frame rate of 14 images/second (i.e. temporal resolution of 72ms). The acquisition of each sequence took about 5 seconds, resulting in 75 coronal oblique images. A previously obtained sagittal slice was used to better determine the orientation of the oblique slice.

Figure 1 provides details concerning the acquired image sequences depicting the oblique acquisition plane and the location of the oral and nasal cavities.

2.3 Data Processing

Data processing consists in three main steps: image segmentation, to identify the oral and nasal cavities and compute their areas for all image frames; audio

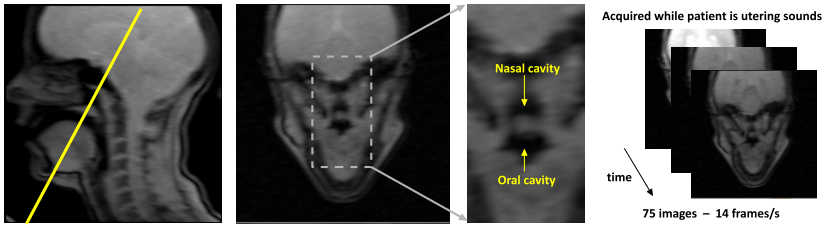


Fig. 1. From left to right: mid-sagittal plane depicting orientation of the oblique plane used during acquisition, sample oblique plane showing the oral and nasal cavities and image sequence details.

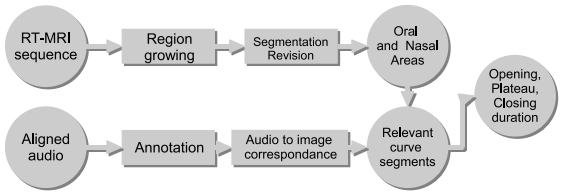


Fig. 2. Data processing pipeline for image and audio data.

annotation, to be able to identify the set of frames corresponding to the nasal vowels; and, finally, analysis of the extracted curve segments. Figure 2 depicts the data processing pipeline.

Since the subject keeps the same head position, throughout the acquisition, the oral and nasal cavity are located in the same region in every image. This was explored in order to speed segmentation of both cavities along the image sequence (75 frames).

A region growing method was used, defining a neighbourhood for each seed including both spatial coordinates and time. This allowed that, with a small number of seeds, each of the cavities could be segmented along the whole image sequence.

Segmentation starts by the oral cavity, where the user has to position a seed. The intensity interval for the region growing (automatically established from the seed neighbourhood) can be tuned for each image sequence but a value around 6% of variation has provided good results overall. The number of seeds needed to fully segment the oral cavity depends on the number of times the cavity closes during the sequence. For the processed sequences only one seed was required to fully segment the oral cavity.

The nasal cavity was segmented using a similar method to the one used for the oral cavity. Since the nasal cavity closes several times along each image sequence, five seeds (in different image frames) were typically needed to perform the segmentation. Further details concerning the segmentation method can be found in Silva et al. [20].

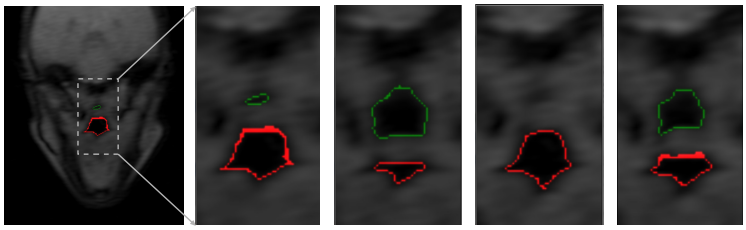


Fig. 3. Different frames depicting the segmented oral and nasal cavities.

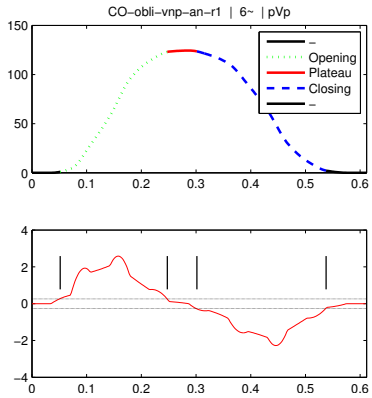


Fig. 4. Top, nasal area variation curve for the [ɐ̃] in [ɐ̃pa] with the different durations depicted in different line types. Bottom, first derivative of nasal area curve depicting the landmarks used to determine duration of velar opening, plateau and closing gestures.

Figure 3 shows sample frames depicting the oral and nasal cavities segmentations. For this particular case the user defined a total of six seed points: one for the oral cavity (in frame 16) and five for the nasal cavity (in frames 0, 25, 28, 37 and 56). Segmentation time was around two minutes for each image sequence.

Audio segmentation was performed manually, using the software tool Praat [21], in order to locate the target sounds. Since the audio is synchronized with the image sequence it is possible to identify the set of image frames corresponding to the production of each vowel in the different prosodic contexts.

Velar opening, plateau and closing durations were computed from the nasal area curves. The first derivative for the nasal area curve segments was computed and passings through 10% of the derivative maximum used as landmarks for the transition between gestures (as illustrated in figure 4).

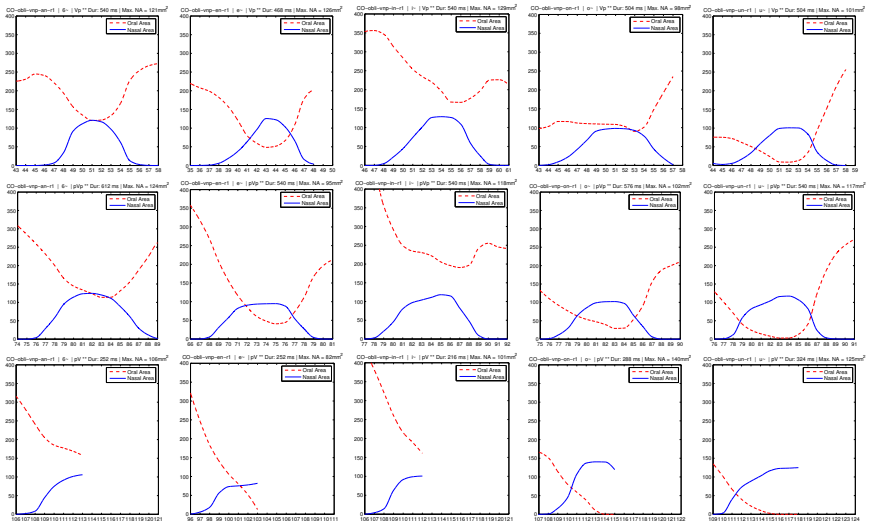


Fig. 5. Time-varying nasal and oral areas for all EP nasal vowels (left to right [ɐ̃], [ẽ], [ĩ], [õ], [ũ]) produced in word-initial (top), word-internal (middle) and word-final (bottom) position.

3 Results

3.1 Nasal/oral area variation over time

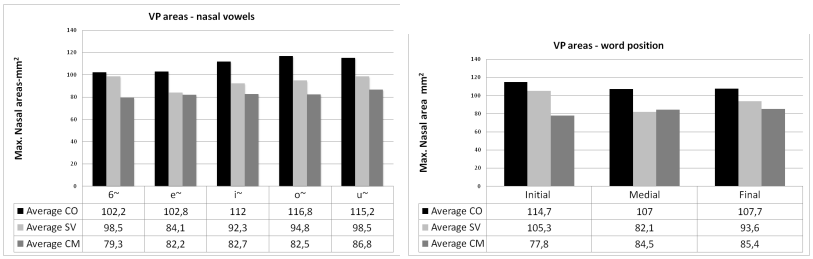
Figure 5 presents, as a representative example of what can be observed for all speakers, nasal and oral areas, over time, for all EP nasal vowels produced by speaker CO in three different word positions flanked by the plosive [p].

In the beginning of the nasal vowel, nasal areas (N) are very small (lower than 25 mm^2) denoting a small velum port opening (VPO). After that, the curves show a gradual increase of the areas until reaching the maximum nasal area (lowest velar position) and, finally, the areas decrease again for the production of [p]. The magnitude of the nasal areas is quite similar for the different vowels (approx. 100 mm^2 for this speaker). This pattern is observed both in word-initial and word-internal position. In word-final position, after reaching the maximum nasal area, the area does not decrease since the velum remains open.

The oral areas (O) follow an inverse trend, decreasing while the nasal area increases. However, the presented curves show differences in oral areas between the five EP vowels. Oral areas are generally higher for [ɐ̃] and [ĩ] than for [õ] and [ũ] in consequence of different tongue configurations among the vowels.

3.2 Velar Port Opening Areas

Figure 6 shows the mean for the maximum nasal areas for each speaker and EP nasal vowel (left) in different word-positions (right).



	ɐ̃	ẽ̞	ĩ	õ̞	ũ
Average CO	102,2	102,8	112	116,8	115,2
Average SV	98,5	84,1	92,3	94,8	98,5
Average CM	79,3	82,2	82,7	82,5	86,8

	Initial	Medial	Final
Average CO	114,7	107	107,7
Average SV	105,3	82,1	93,6
Average CM	77,8	84,5	85,4

Fig. 6. Left, average of the maximum nasal areas for each of the EP nasal vowels produced by the 3 speakers in all positions considered in this study; right, average of the maximum nasal areas for each of the three word positions.

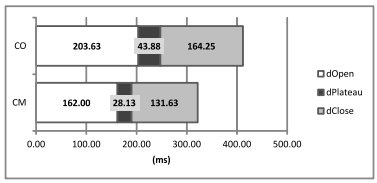


Fig. 7. Mean duration of velar opening, plateau and closing for speakers CO and CM.

Comparing the results obtained, differences in the nasal areas between the three subjects can be observed. Subject CO presents the highest nasal areas (between 102,2 mm^2 for [ɐ̃] and 116.8 mm^2 for [õ̞]), followed by subject SV (between 84,1 mm^2 for [ẽ̞] and 98.5 mm^2 for both [ɐ̃] and [ũ]) and, finally, subject CM (between 79.3 mm^2 for [ɐ̃] and 86.8 mm^2 for [ũ]). This reflects anatomical differences between subjects, though not very accentuated, as they are all female with similar biotype.

Figure 6, on the left, suggests only slight differences in nasal areas among vowels, for all speakers. The observed differences in VPO do not seem to be related with differences in tongue position of the vowels.

As illustrated in Figure 6, on the right, VPO areas of the nasal vowels produced in different word positions are rather similar to each other, at least for CO and CM. Subject SV reveals a difference of 23 mm^2, between initial and medial word position.

3.3 Duration of velum movements

Figure 7 shows the mean durations (in ms) of velum opening (d_{Open}), plateau ($d_{Plateau}$) and closing (d_{Close}) gestures of EP nasal vowels computed for two of the speakers, CO and CM.

Mean total duration of velar opening - closure gesture is 437 ms for speaker CO and 333 ms for speaker CM. As shown in figure 7, mean d_{Open}, $d_{Plateau}$ and d_{Close} are respectively 203.63 ms/43.88 ms/164.25 ms for speaker CO and 162.00

Speaker	[ɐ̃]	[ẽ]	[ĩ]	[õ]	[ũ]
CO	0.47	0.66	0.39	0.60	0.97
SV	0.30	0.30	0.35	0.76	0.99
CM	0.39	0.47	0.34	0.96	0.99

Table 1. Mean values of observed maximum oral-nasal areas proportion, P_{NO}, by speaker and vowel.

ms/28.13 ms/131.63 ms for speaker CM. For both speakers, opening gesture is longer than closing movement.

The d_{Open} variation pattern, for each vowel, is similar for the two analysed speakers: short d_{Open} for [ũ], followed by [ɐ̃], [ĩ] and [õ], with similar duration values, and finally [ẽ], presenting the longest duration. For all vowels, d_{Open} is longer than d_{Close}, except for [ũ].

3.4 Nasal-oral areas proportion

The nasal-oral proportion, $P_{NO} = N/(N + O)$, was computed, as well as its variation over time. As expected, the variation of the P_{NO} over time follows a similar pattern to the one observed for the nasal areas. At the beginning of the nasal vowels, P_{NO} is nearly zero, increasing over vowel production and decreasing again for [p] articulation.

However, the magnitude of the P_{NO} differs between nasal vowels. Since the nasal areas are similar among vowels, as previously shown, this variation is mostly due to differences in oral areas.

The mean maximum values of P_{NO} by speaker and vowel are shown in Table 1. For all the speakers, the P_{NO} is higher for back vowels ([õ] and [ũ]) than for the more fronted ones ([ɐ̃], [ẽ] and [ĩ]).

4 Conclusions

The results obtained in this study corroborate previous accounts for EP, namely that nasality in Portuguese is typically incremental over the vowel, with a movement from oral to nasal [14, 10]. This trend was consistently observed for the three speakers, in all vowels and word positions. As expected, in word-internal position, nasal vowel is partly produced as an oral vowel (i.e. nasal areas are very small), since the nasal vowel follows a plosive, that only allows slight velopharyngeal port openings [22]. However, in initial position, small nasal areas were also observed, that reflect a closing movement of the velum, when it would be predicted a continuity of the velum opening from the rest position.

Moreover, nasal area values suggest that the magnitude of the velopharyngeal port opening is similar between the different vowels. These findings are in agreement with previous EMA data, obtained by Oliveira [9], for one speaker, that indicated there were no significant differences in velum height between nasal vowels. Although velum height and VPO are different measures, they are correlated [23].

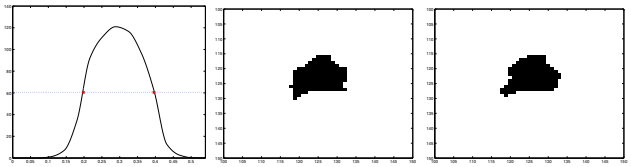


Fig. 8. Velar port configuration for corresponding areas (left) during opening (centre) and closing (right) movements.

The mean durations of velum opening, plateau and closing are, in general, consistent with previous EMA data [13, 10]. The opening movement is longer (25–35 ms) than the closing movement, for both speakers and vowels, excluding the high back nasal vowel ([ũ]).

In sum, oblique RT-MRI slices at the velopharyngeal port provided new articulatory data on time varying nasal areas, that complement and clarify previous results obtained with EMA, allowing for a better characterization of velum movement in the production of nasal vowels.

The available image data might also allow the study of velar port configuration (shape) over time (figure 8), another important feature for proper modelling of EP nasal vowels. Furthermore, temporal organization of gestures will be addressed, through the analysis of real-time image sequences acquired both at the velum and lips.

Acknowledgements

This research was partially funded by FEDER through the Operational Program Competitiveness factors (COMPETE) and by National Funds through Foundation for Science and Technology (FCT) in the context of the Project HERON II (PTDC/EEA-PLP/098298/2008). The first author acknowledges the PhD grant from FCT (SFRH/BD/65183/2009). The authors thank Professor Miguel Castelo-Branco (IBILI Director) and the speakers involved in the study.

References

1. Amelot, A., Crevier-Buchman, L., Maeda, S.: Observations of velopharyngeal closure mechanism in horizontal and lateral direction from fiberscopic data. In: ICPhS, Barcelona (2003) 3021–3024
2. Zerling, J.P.: Phénomenes de nasalité et de nasalization vocaliques: Étude cinéradiographique pour deux locuteurs. Travaux de l'Institut de Phonétique de Strasbourg **16** (1984) 241–266
3. Amelot, A., Rossato, S.: Velar movements for two French speakers. In: ICPhS, Saarbrücken (2007)
4. Engwall, O., Delvaux, V., Metens, T.: Interspeaker variation in the articulation of French nasal vowels. In: In Proceedings of the Seventh International Seminar on Speech Production, Ubatuba, Sao Paolo, Brazil (2006) 3–10

5. Delvaux, V., Demolin, D., Harmegnies, B., Soquet, A.: The aerodynamics of nasalization in French. Journal of Phonetics **36** (2008) 578–606
6. Maeda, S.: Acoustics of vowel nasalization and articulatory shifts in French nasal vowels. In Huffman, M.K., Krakow, R.A., eds.: Nasals, Nasalization, and the Velum. Phonetics and Phonology (vol. 5). Academic Press Inc. (1993) 147–167
7. Rossato, S., Teixeira, A., Ferreira, L.: Les nasales du portugais et du français: une étude comparative sur les données EMMA. In: Journées d'Études sur la Parole (JEP), Dinard, France (2006) 143–146
8. Oliveira, C., Teixeira, A.: On gestures timing in European Portuguese nasals. In: ICPhS, Saarbrücken (2007)
9. Oliveira, C.: Do grafema ao gesto: Contributos Linguísticos para um Sistema de Síntese de base articulatória. Dissertação de doutoramento, Universidade de Aveiro, Aveiro (2009)
10. Oliveira, C., Martins, P., Teixeira, A.: Speech rate effects on European Portuguese nasal vowels. In: InterSpeech, Brighton (2009)
11. Martins, P., Carbone, I., Silva, A., Teixeira, A.: An MRI study of European Portuguese nasals. In: Interspeech, Antuérpia (2007)
12. Martins, P., Carbone, I., Pinto, A., Silva, A., Teixeira, A.: European Portuguese MRI based speech production studies. Speech Communication **50**(11-12) (2008) 925–952
13. Teixeira, A., Vaz, F.: European Portuguese nasal vowels: An EMMA study. (2001)
14. Lacerda, A., Head, B.: Análise de sons nasais e sons nasalizados do português. Revista do Laboratório de Fonética Experimental de Coimbra **6** (1966) 5–70
15. Teixeira, A., Vaz, F., Príncipe, J.: Influence of dynamics in the perceived naturalness of Portuguese nasal vowels. In Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., Bailey, A.C., eds.: ICPhS, San Francisco, University of California, Berkeley (1999)
16. Demolin, D., Delvaux, V., Metens, T., Soquet, A.: Determination of velum opening for French nasal vowels by magnetic resonance imaging. Journal of Voice **17**(4) (2003) 454–467
17. Teixeira, A., Martinez, R., Silva, L.N., Jesus, L., Príncipe, J., Vaz, F.: Simulation of human speech production applied to the study and synthesis of European Portuguese. EURASIP Journal on Applied Signal Processing **9** (2005) 1435–1448
18. Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R.: Real-Time MRI for Portuguese. Database, methods and applications. In: PROPOR, Coimbra (2012)
19. Martins, P., Carbone, I., Pinto, A., Silva, A., Teixeira, A.: European Portuguese MRI based speech production studies. Speech Communication **50**(11–12) (2008) 925–952
20. Silva, S., Martins, P., Oliveira, C., Silva, A., Teixeira, A.: Segmentation and analysis of the oral and nasal cavities from MR time sequences. In: Proc. 8th International Conference on Image Analysis and Recognition (ICIAR 2012), LNCS vol. 7325. (2012) 214–221
21. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [computer program] (2012)
22. Warren, D.W., Dalston, R.M., Mayo, R.: Aerodynamics of nasalization. In Huffman, M.K., Krakow, R.A., eds.: Nasals, Nasalization, and the Velum. Phonetics and Phonology (vol. 5). Academic Press Inc. (1993) 119–146
23. Serrurier, A., Badin, P.: A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. J. Acoust. Soc. Am. **123**(4) (2008) 2335–2355

Subjective Evaluation of a High Quality Text-to-Speech System for Argentine Spanish

Jorge Gurlekian, Christian Cossio-Mercado, Humberto Torres and M. Elena Vaccari

Laboratorio de Investigaciones Sensoriales (LIS), INIGEM CONICET-UBA
Av. Córdoba 2351 Piso 9 Sala 2, Buenos Aires, Argentina
{jag,htorres,mvaccari}@fmed.uba.ar,cgcossio@rucatech.com.ar

Abstract. This work summarizes the perceptual evaluation of our recently developed text-to-speech system (S1), based on unit concatenation. We compare it with two commercially available systems (S2 and S3) using three different evaluation methods. One is the P.85 recommendation by the International Telecommunication Union (ITU), the second method, called Syntactically Unexpected Sentences (SUS), is known to be the most strict for intelligibility evaluation, and the third is the Mean Opinion Score (MOS) scale. Results of ITU test showed better quality and intelligibility responses for system S2. General quality evaluated by MOS and intelligibility evaluated by SUS presented no appreciable differences between S1 and S2. It is concluded that high quality performance is related to a complete intonation modeling of different type of phrase length and styles. S1 has high intelligibility and quality for general information sentences but presented lower scores for specific tasks as proposed in ITU tests, where short phrases coverage should be introduced in the intonational modeling of our system.

Keywords: TTS Evaluation, Intelligibility, Speech Quality, ITU P.85, Mean Opinion Score, MOS, Syntactically Unexpected Sentences, SUS

1 Introduction

Our laboratory has proposed to develop a synthesis system where the primary objective is to achieve natural quality, additionally to high intelligibility. For this goal we have analyzed Buenos Aires Spanish prosody [1–5] and proposed a way to incorporate this knowledge into a TTS system. A development project is in progress, based on segment concatenation to ensure good intelligibility and to allow the incorporation of prosody target contours from text [4].

There are several trade-offs related to the quality goal during the design of a TTS concatenative system. The main ones are unit and database sizes. Concatenation units could be words, syllables or diphones. As the unit gets smaller the less the quantity of them that are needed, though quality drops off. A possible solution is to combine standard diphones units with polyphones [9]. Moreover, in order to produce high quality voices database size must have a high coverage of units related to different prosodic targets, not just in different phonemic contexts.

In order to produce a new high quality voice a totally new database must be recorded with more than 1000 sentences. For example, to obtain high quality in a tonal language like Chinese Mandarin, 12000 sentences should be used; even this amount only solves 64% of the tonal mandarin variations [10] and when using 1200 sentences, evaluation on a 5-points MOS scale showed that quality decreased significantly from 2.93 to 2.36. As for Spanish, there are several academic and commercially available TTS systems that provide a number of voices for Latin American varieties, and even Argentine.

1.1 Evaluated Systems

We compared our system (S1) with other well known solutions: Nuance TTS (S2) and AT&T Natural Voices (S3). All three systems are based on unit concatenation. Our voice was created using an Argentinian female speaker, while samples for Nuance [15] were generated using Paulina

(female, mexican) and Rosa (female, neutral Latin American) for AT&T [14], since these voices were the most similar to ours. All audios were normalised in gain and sampling rate (8 kHz).

Our diphone database was created from 1593 sentences, taken mostly from newspapers, that guaranteed unit coverage for a total of 85 min, 14000 words and 68300 units. By working with a professional speaker, both good articulations and variations were produced, with one to three tonal phrases, but sometimes tonal accents were located on syllables not lexically accented, though this issue was postponed for posterior analysis.

For intonation contour prediction from text we utilized the Fujisaki model and CART's for their parameter estimation. Neural Networks were employed for segmental duration prediction. Finally, automatic diphone selection according to variable costs allowed us to achieve real time performance.

1.2 Methods of Evaluation

We used three different methods for subjective evaluation:
The first is the P.85 recommendation [6], that was introduced in 1994 by the International Telecommunication Union (ITU) in order to assess subjective performance for overall quality of speech voice output devices. It considered that signal manipulation and processing, like digital compression or concatenation, always implies degradation of quality. The recommendation covers voice servers that are available at the commuted public network, and pre-recorded signals and synthetic voices are included. The term *synthetic voice* includes those completely artificial build up by parametric formant synthesizers [9, 13] as well as the resulting of unit concatenation of variable length [4]. Tests emerging from ITU standards are aimed at applications highly dependent on the task using long sentences with short phrases, where there are a fixed carrier text and others variable segments such as names and numbers.
The Semantically Unpredictable Sentences (SUS) test [8] is focused on evaluating the intelligibility of systems. SUS sentences are syntactically correct but have no meaning or present a very low word predictability.
The third method is the standard Mean Opinion Score (MOS) [7], where listeners use a fixed scale from 1 to 10 to evaluate overall quality. This is a general purpose test and provides average scores for natural and artificial speech.

Recently Sitayev et al. [11] compared ITU, SUS and MOS tests for English. They claim that ITU test does not have the same discriminative power as SUS, which is the most strict and useful for intelligibility testing. Regarding the comparison with MOS, results from ITU test which uses a 5-point MOS compared to the standard 10-point MOS test show a different order of quality for the systems evaluated. They conclude that the type of prosodic structure of the sentence plays an important role and worsen some particular problems of the systems. For example, in short sentences the *pleasantness* of a voice could dominate the *quality* judgment, while in long sentences *speaking rate* could be the main factor.

Finally, most works (e.g., [11, 12]) show that intelligibility and quality have a weak correlation. The same occurs for acceptability and preference. Acceptability is related with overall quality but the pleasantness of a voice drives the answers to the preference decision.

2 Experimental Design

A total of twenty listeners, 23–45 years old, participated in the experiment. Ten of them were Speech Language Therapists (SLT) both professionals and students of the University of Buenos Aires. The other ten were non-expert (non-SLT) subjects without auditory impairments. All listeners were paid for their work. They were contacted by a university professor who sent mails with instructions and stimuli[1]. Answers were received in a text file for automatic processing. In Annex is available the full list of text used for each of the parts of the test.

[1] All the samples for our system are available at `http://www.lis.secyt.gov.ar/papers/TTSEvaluation12/LIS-TTS_stimuli.zip`

2.1 ITU Test

In total 27 sentences were created from 9 different texts. They have between 20 and 25 words, and up to 10 short melodic phrases. Listeners could hear stimuli as often as required to make judgments of intelligibility and quality. The sentences have a fixed portion specific to the task and a variable part that changes in each presentation producing several melodic groups. Texts were designed for three tasks: telephone sales, flight information and payment services.

Subjects listened to 9 different sentences (3 from each system) in each of two sessions, giving a total of 18 audios per listener. The evaluation on general aspects of intelligibility (i) was done in the first session while quality evaluation (q) was presented in the second session. Below is a summary of the instructions given to the subjects, the scales used and in which part they appeared.

Word Recognition (i) You have to write down [name, product, features, code, price and time], [company, country, flight number, departure time, terminal and boarding gate] or [name, company, month, amount and branch] (according to each task).

Overall Impression (q) How is the quality of what you heard: Excellent; Good; Fair; Poor; Bad.

Listening Effort (i) How would you describe the effort made to understand the message: None; Low; Moderate; High; Not understood with any effort.

Comprehension Problems (i) There are words difficult to understand: Never; Rarely; Occasionally; Often; All the time.

Articulation (i) The sounds are distinguishable: Yes, clear; Yes, clear enough; Fairly clear; Not very clear; Not at all.

Pronunciation (q) Did you note any anomalies in pronunciation: No; Yes, but not annoying; Yes, slightly annoying; Yes, annoying; Yes, very annoying.

Speaking Rate (q) The average speed was: Much faster than preferred; Faster than preferred; Preferred; Slower than preferred; Much slower than preferred.

Voice Pleasantness (q) How would you describe the voice: Very pleasant; Pleasant; Fair; Unpleasant; Very unpleasant.

Acceptance (i, q) Do you think that this voice could be used in an information service by telephone: Yes; No.

2.2 SUS Test

In total 50 semantically nonsensical texts were designed with correct syntactic structure, a length of 6 to 10 words, and containing one or two melodic phrases. For each listener we employed 45 randomized sentences, 15 for each system, without repetitions. Instruction to subjects was: "Write down each word that you heard in the sentence. There will be no replaying, so be alert".

2.3 MOS Test

A total of 60 texts were synthesized, each with 10 to 20 words in two or three long melodic phrases. Subjects listened to 15 sentences, 5 from each system, with the possibility of repetitions from one system to another, but without repetitions for the same system.

3 Results

3.1 ITU Test

As seen in Figure 1, ITU Intelligibility test show similar results for the threes systems, while S3 presents lower performance according to non-experts. The order of the systems with respect to *listening effort*, *comprehension* and *articulation* is S2, S1 and S3 for both groups of listeners.

According to the results for the ITU Quality session, shown in Figure 2, S2 is clearly on top of the responses for all categories. S1 and S3 reached the same performance in *overall* and

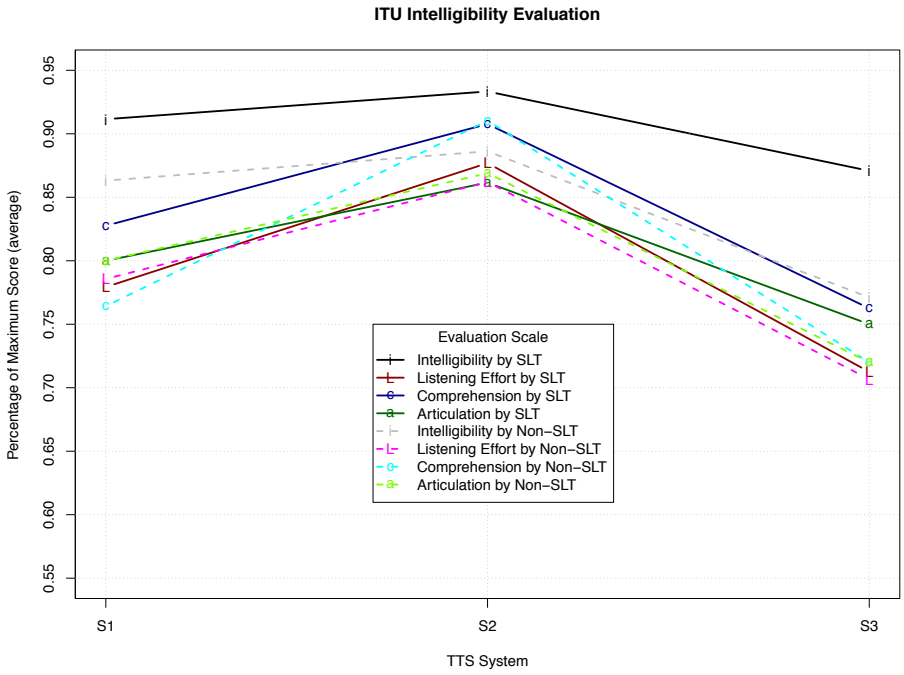


Fig. 1. Average responses of the ITU Intelligibility test for the three systems and two groups of listeners

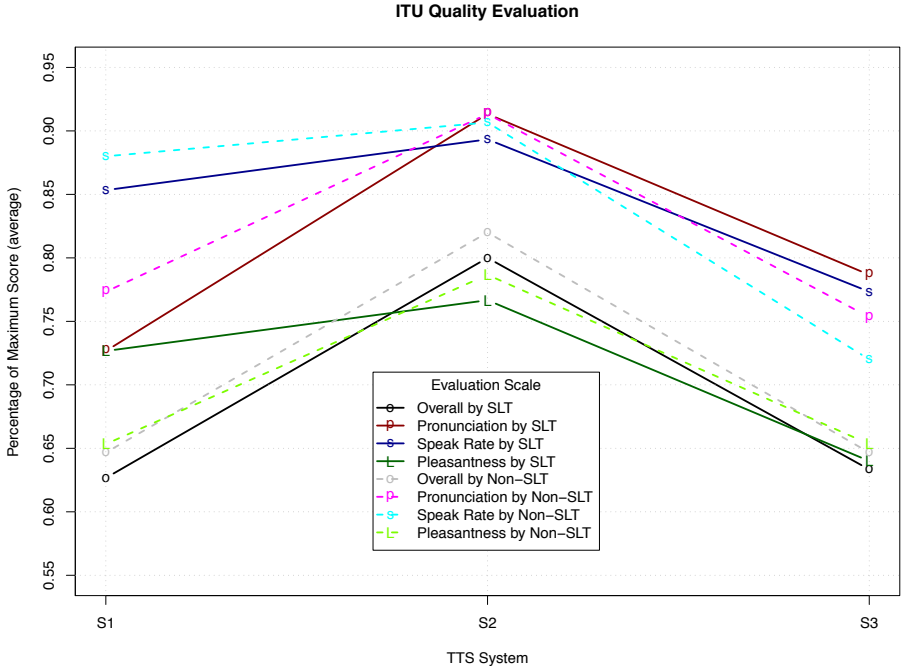


Fig. 2. Average responses for the quality ITU test for the three systems and two groups of listeners

pronunciation scales, both behind S2. Regarding *speaking rate*, S1 and S2 are close and S3 is below them. There is no agreement between experts and non-experts for *pleasantness* of S1.
Responses for *acceptability* by group are presented in Figure 3. Overall, S2 attained to 83% of acceptance, S1 62%, and S3 reached 49%.

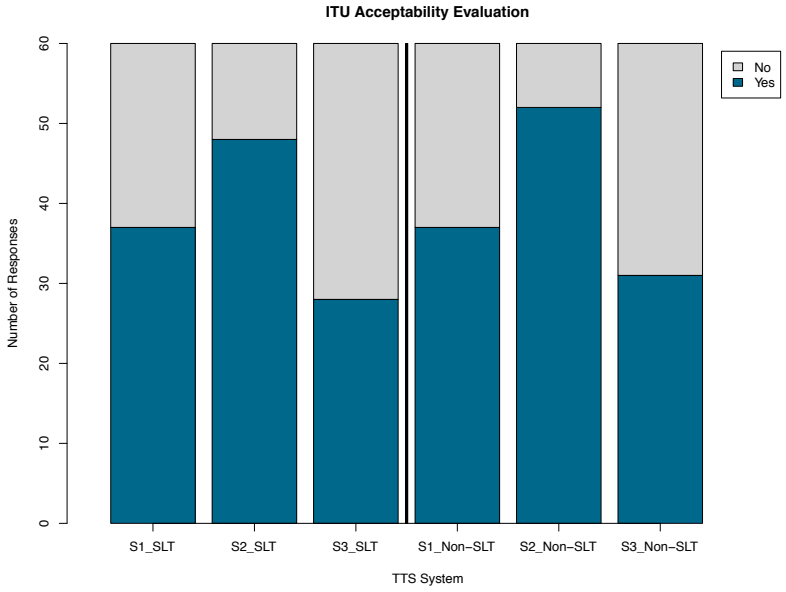


Fig. 3. Responses of ITU Acceptability test for the three systems and two groups of listeners

3.2 SUS Test

According to the percentage of correct words per phrase of SUS test, systems are ordered as S1, S2 and S3 for non-experts. For experts, responses show an equivalence between S1 and S2.

Regarding the number of correct phrases, S1 shows better results than S2 for non-experts, while S2 outperforms S1 according to experts. None of these differences are statiscally significant in a proportion test, but differences of both systems with respect to S3 do (respectively: $\chi^2 = 50.5457$, df = 1, $p < 0.00001$; $\chi^2 = 58.0703$, df = 1, $p < 0.00001$).

Analysis of variance for the SUS test shows that significant differences exist in the scores of the three systems (S1–S2, F = 0.5752, num df = 314, denom df = 314, $p \ll 0.001$; S1–S3, F = 0.3005, num df = 314, denom df = 314, $p \ll 0.001$, S2–S3, F = 0.5225, num df = 314, denom df = 314, $p \ll 0.001$), and that S1 has the smallest variance of the three systems. The post-hoc analysis indicates no significant differences between the means of systems 1 and 2, and both differ significantly from S3 (respectively: t = 8.1123, df = 487.093, $p < 0.00001$; t = 7.0721, df = 571.753, $p < 0.00001$).

3.3 MOS Test

Results for the MOS test indicates an order for *overall quality* of S2, S1 and S3 for non-experts, as seen in Figures 6 and 7. For expert listeners S1 and S2 responses are even. Analysis of variance for the 10 point MOS test indicated that no significant differences exist among the scores of subjects for the sentences of the three systems. The post-hoc analysis indicates no significant differences between means of systems 1 and 2, while both significantly differ from S3 (respectively: t = 6.1381, df = 207.015, $p < 0.00001$; t = 7.323, df = 207.669, $p < 0.00001$).

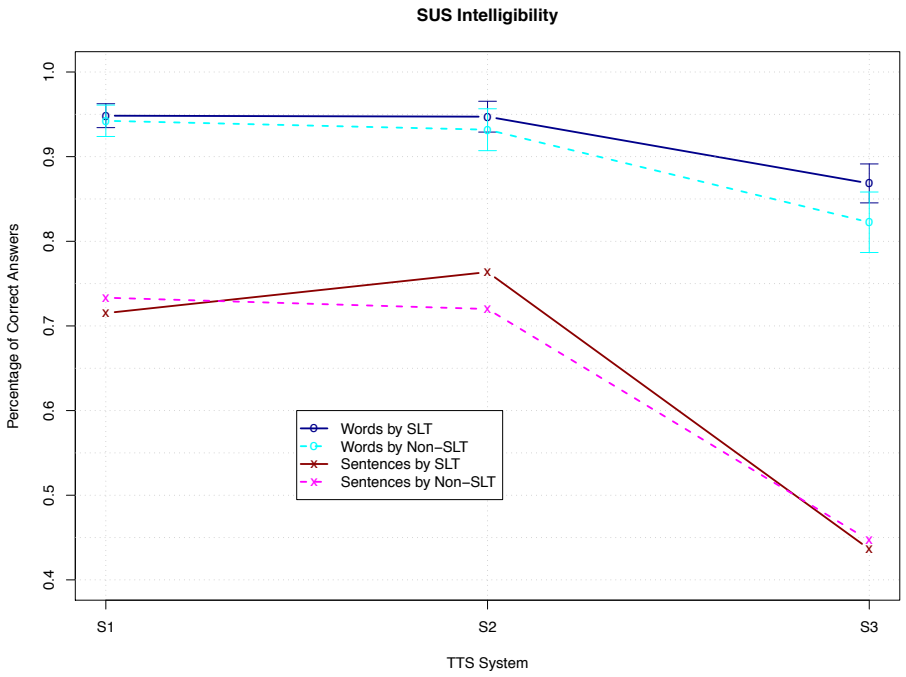


Fig. 4. Averaged percentage of correct words per phrase and porcentage of correct phrases from ITU test, for the three systems and two groups of listeners

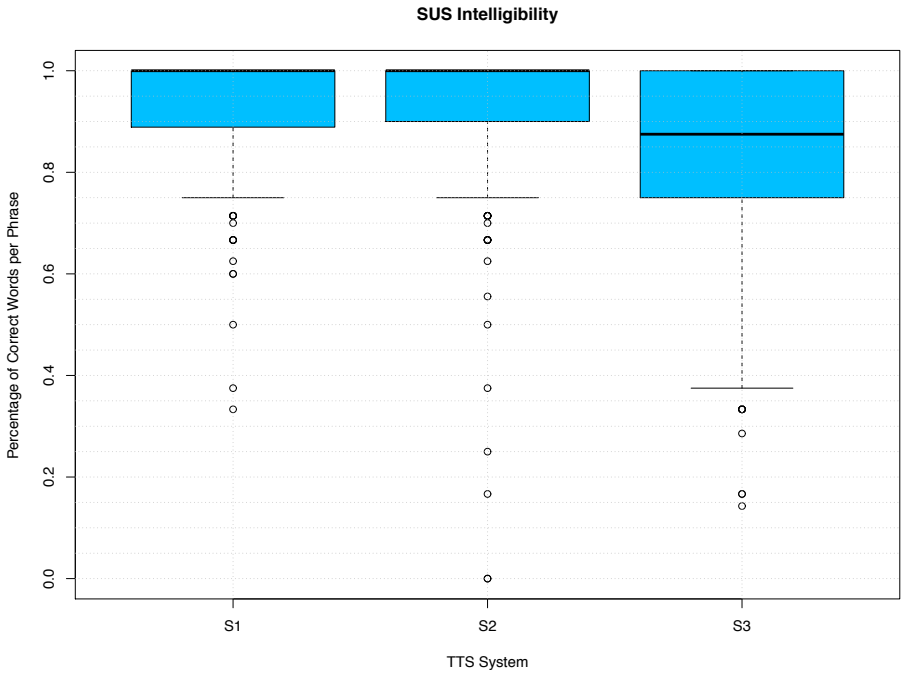


Fig. 5. SUS evaluation of correct words for the three systems evaluated over 20 listeners

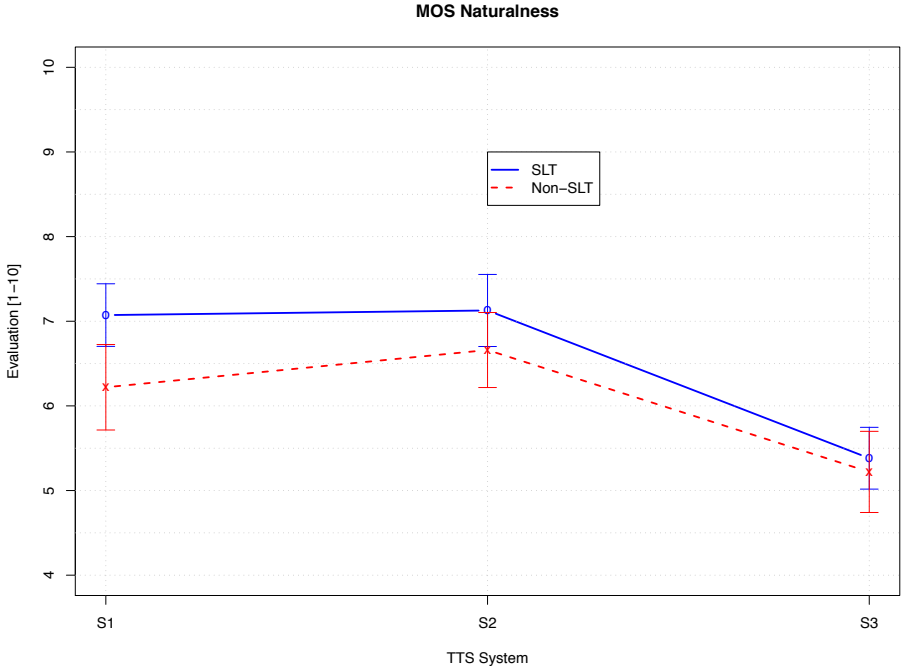


Fig. 6. Average responses of MOS evaluation by the two groups of listeners

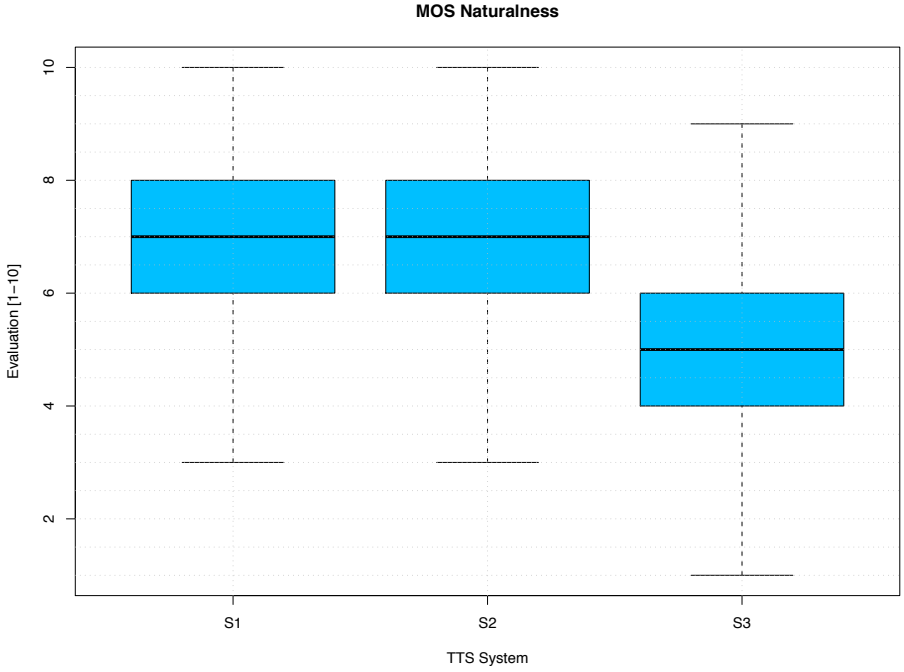


Fig. 7. MOS evaluations for the three systems tested over 20 listeners

4 Discussion

Intelligibility results from the ITU test show that S1 and S2 systems reach similar response levels. S2 leads with 91%, S1 has 89% and S3 attained the lowest value with 82%. Intelligibility measured with the SUS test showed a similar performance: S1 at the first place, with 95%, very close S2, with 94%, while S3 received 85% of score.
In both tests no significant differences were found between systems S1 and S2 and between experts and non-experts. SUS test revealed more significant statistical differences of S1 and S2 compared to S3. We found no contradictions for intelligibility evaluation between ITU test and SUS test. Besides, *listening effort*, *comprehension* and *articulation* are highly correlated (0.74, t = 13.3538, df = 141, $p \ll 0.001$) and could be collapsed into a single category with *pronunciation*.

There is a high correlation between *overall quality* and *preference* in ITU Quality test (0.7, t = 13.045, df = 178, $p \ll 0.001$). This is even higher for non-experts, what is understandable given that SLT's are trained to analyze fine attributes of speech.

Quality results obtained from the 5-points MOS test of ITU test show a clear order of quality of S2, S1 and S3. Nevertheless, when using 10-points MOS there were identical results for S1 and S2 systems, but responses to S1 presented a smaller variance. We argue that this difference could be due to the type of sentences used in each test. MOS test with 5 points included sentences with more than five melodic groups, signed by forced punctuation marks, while MOS test with 10 points, even with similar quantity of words, employed only one to three melodic groups.

System 1 was created using a database containing sentences with a standard syntactic structure with few melodic groups, and the intonation model was based in those sentences. Thus, our system is more appropriate for the stimuli employed at 10 point MOS test. It is not enough to have a good coverage of phonetic units but also to support different melodic phrases as those used in ITU tests. That is to say that a forced short intonation contour as needed for the type of sentences employed in ITU tests was not correctly modelated. In this way a system model that uses fixed inflexions causes less perturbation to the listeners than one that presents wrong inflexions.

System 2 received higher scores in quality judgments for all scales. System 1 received significant lower scores compared with S2 for *overall quality* and *pronunciation* (t = -5.6318, df = 114.739, $p \ll 0.0001$; t = -5.5743, df = 108.474, $p \ll 0.0001$), but no significative differences were found in *speak rate* and *pleasantness*. System 3 scored significatively lower in all scales relative to S2 and significatively lower in *speak rate* and *pleasantness* than S1. Pronunciation issues could be related to a limited number of diphones and/or a lack of units for proper names and foreign words.

5 Conclusions

System under evaluation (S1) revealed high intelligibility as evidenced by the results obtained in ITU and SUS tests. Quality for sentences with few melodic phrases is also high and comparable with one of the best known commercial systems (S2). There were not significant differences between expert/non-expert listeners for all the tested scales.

Results for specific tasks as proposed by the ITU test, composed by several small melodic groups indicated by forced marks, revealed lower *quality* and *pronunciation* scores than the best system tested.

In order to improve performance we conclude that variable length phrases must be included in our system S1. As is referenced in the description of the compared systems for English, the use of a multilingual database could better solve foreign names or proper nouns. Besides, the use of non uniform units in sytems 2 and 3 versus fixed diphone units used in S1 should also avoid some slight concatenation disruptions.

6 Acknowledgments

The development of the TTS system was supported by the Ministry of Science and Technology and Productive Innovation of Argentina (MinCyT), as part of project PAE-PID-2007-00094 "High quality text to speech conversion".

References

1. Colantoni, L. and Gurlekian, J.A.: Convergence and Intonation: Historical Evidence from Buenos Aires Spanish. Bilingualism: Language and Cognition. Penn State Univ. Vol.7, Nr.2, pp. 107–119, 2004.
2. Torres, H., Mixdorff, H., Gurlekian, J.A., and Pfitzinger, H.R.: Two New Estimation Methods for a Superpositional Intonation Model. Proc. of Interspeech 2010. ISCA, 2010.
3. Torres, H.M. and Gurlekian, J.A.: Parameter Estimation and Prediction from Text for a Superpositional Intonation Model. Elektronische Sprachsignalverarbeitung. Dresden. September 2009. pp. 238–248. 40 Jahre IAS, 20 Jahre ESSV. 2009.
4. Torres, H.: Generación Automática de Prosodia para un Sistema de Conversión de Texto a Habla. PhD Thesis. UBA, 2008.
5. Torres, H. and Gurlekian J.A.: Automatic Determination of Phrase Breaks for Argentine Spanish. Speech Prosody, Nara, Japan. pp. 553–556, Mar. 2004.
6. ITU-T Recomendation P.85. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. ITU, 1994.
7. ITU-T Rec. P.800. Methods for Subjective Determination of Transmission Quality. ITU, 1996.
8. Benoît, C., Grice, M. and Hazan, V.: The SUS Test. A method for the assessment of TTS synthesis intelligibility. In: Speech Communication, vol. 18, 381–392, 1996.
9. Torres, H.M. and Gurlekian J.A.: Acoustic Speech Unit Segmentation for Concatenative Synthesis. Computer Speech and Language, Elsevier, Vol 22, No. 2, 2007.
10. Chu, M. and Peng, H.: An Objective Measure for Estimating MOS of Synthesized Speech. Eurospeech 2001. Scandinavia. 2001.
11. Sityaev, D., Knill, K. and Burrows, T.: Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems, in: Proc. of Interspeech 2006. Pittsburgh PA, 2006.
12. Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A., Swerts., M. and Terken, J.: Evaluation of Speech Synthesis Systems for Dutch in Telecommunication Applications, In Proc. of 3rd ESCA Workshop on Speech Synthesis, 1998.
13. Klatt, D.H. and Klatt, L.C.: Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. Journal of the Acoustical Society of America 87, pp. 820–857, 1990.
14. AT&T TTS site, `http://www2.research.att.com/~ttsweb/tts/`.
15. Nuance TTS site, `http://www.nuance.com/vocalizer5/`.

Annex

ITU Test Sentence List

Task 1

Señor Di Fiori: el televisor con pantalla 3D, de 52 cm., código: tres uno seis, de 42000$, se le enviará en una semana.

Señorita Pérez: la cafetera exprés de acero inoxidable, triple pocillo, con código 5734, de 1235$, se le enviará en 3 días hábiles.

Señora Embe: la máquina de coser Sínger, de puntada invisible, código 4197, de 1399$, se le enviará en 3 horas.

Task 2

Atención. Aerolíneas Argentinas anuncia que el número de vuelo, seis cuatro tres, con destino a Italia, saldrá a las 12:10hs, de la terminal B. Puerta de embarque 5.

Atención. Américan Erlayns anuncia que el número de vuelo, ocho seis tres, con destino a Méjico, saldrá a las 8:30hs, de la terminal C. Puerta de embarque 14.

Atención. Yapán Erlayns anuncia que el número de vuelo, 1594, con destino a Tokio, saldrá a las 18hs, de la terminal A. Puerta de embarque 12.

Task 3

Estimado Jorge Pérez. La factura de Telatel del mes de enero, por 214$, está pendiente de pago. Le pedimos que se acerque a la sucursal Flores para regularizar la situación. Gracias.

Estimado Joaquín González. La factura de Tututel del mes de febrero, por 453$, está pendiente de pago. Le pedimos que se acerque a la sucursal Palermo para regularizar la situación. Gracias.

Estimada Susana Fachineli. La factura de Tatatel del mes de marzo, por 784$, está pendiente de pago. Le pedimos que se acerque a la sucursal Belgrano, para regularizar la situación. Gracias.

SUS Test Sentence List

El viento dulce armó un libro de panqueques
El insecto francés conduce el elevador
El avión pintaba los días de melón cocido
El pincel construyó algunos océanos exitosos
A las dos se suben los meses colorados
Los meses cocinan zapatos de bambú
Podemos atrasar el camión con chocolate de acero
La bicicleta contiene cinco elefantes voladores
El cuadro es una creación con harina de caqui
Sin barco nos dirigimos con presión a la peluca
El chancho no escribe las pinturas de flan con agua
Estamos por comer la galaxia sin honor
Los militares extranjeros beben un plato de cemento
Las flores bebieron un ascensor divertido
El viento amargo armó un libro de maní
El caballo de detergente conduce la heladera
La piraña cantó con el mate de la biblioteca
La perra quería dominar al café del mar
Estudiaba el tribunal con pelos del río
El cantante toma el corcho con cera paraguaya
Las lluvias llegan a carcajadas de un mantel
El camión danzaba con luces de miel rugosa
El libro chillón cantaba crema de zapatos
Ellos corrían con signos y choripanes
Milanesas con aire otoñal de anillo
Los salames escribían melodías sabrosas
Saben de su amor por el almíbar con patas
El amor de la fruta de zapatos y la maceta
Se vacunan con semillas de llave inglesa
La receta es con fiambre y bulones de plástico
Las hojas del perchero querían contar
Pintaban pieles con acero quirúrgico
El humo de la estación cantaba muy feliz
El micrófono de la papa quedaba en Brasil
La sal endulzaba la puerta de madera
Todos sabían que la migraña bailaba con Violeta
El melón ya sabía la verdad del hielo
Cantó con la botella del tubo fluorescente
El oso panda estudiaba el rollo de cocina
La cama corría maratón con el balero
Tuvieron sueños de miel con madera plastificada
La bolsa mostraba los ojos con alquitrán
No querían condimentar las uvas de cemento
Vamos a comer empanadas de neumáticos
La nieve se enamoró de un enano de polenta
Ganaron una casa de humo celeste
Exprimieron tubos de acero con azúcar
Mezclaron arroz crudo y alambres de papel
Le gustan las películas con pecas y alfajor
La piedra tocaba la guitarra con las ramas

MOS Test Sentence List

De cada seis pacientes que se van a hacer un estudio, sólo atienden a dos
El plantel volvió a entrenar ayer en el Parque General San Martín
La autora convirtió el material en un éxito de taquilla global
Si hubo un responsable dentro del gobierno, será sancionado
En las próximas horas, habría más cambios en el Gabinete
Y la principal preocupación de los jugadores es zafar de la Promoción
Este fin de semana, quedó como único puntero del torneo local
Hay gustos, que se pagan carísimo
Este no es el momento adecuado para discutir
Si alguien tiene pruebas, que las presente ante la Justicia
Los dirigentes del gremio, confían en que la Presidente los recibirá
Parece que sabían los movimientos de la familia
El sector de informática, es el nuevo generador de empleo del país
La propuesta es refinanciar, y así salir de la depresión económica
La segunda semana fue totalmente exitosa
Este es un partido clave en la batalla por evitar la Promoción
Lo que ocurrió aquí, es algo muy terrible
Esperamos que resulte según lo previsto
El resto de la escena, se completa en forma virtual
En los próximos años, se estima que el clima recrudecerá lentamente

Design of a message verification tool to be implemented in CALL systems

Igor Odriozola[1], Inma Hernáez[1], Eva Navas[1]

[1] Aholab signal processing laboratory, Urkixo zumarkalea z/g, Bilbao, Basque Country
{igor,inma,eva}@aholab.ehu.es

Abstract. This paper presents basic research for the development of a message verification system in Basque to be implemented in CALL (*Computer-Assisted Language Learning*) applications. The system aims to verify a sentence uttered by the user in real time, word by word, in order to display the verified word as soon as it is detected. First a decision threshold for the PS (*Phoneme Score*) is calculated by means of inserting artificial errors in the system. Then, the structure of the ASR internal lattice is described, which includes a phoneme loop between words to absorb the effects of unexpected speech. A last experiment has been carried out to check the behavior of the whole system simulating the insertion of an erroneous extra word by the user. The results of the experiments show that the proposed system is suitable for message verification tasks.

Keywords: CALL systems, utterance verification (UV), message verification, PS scores, L2 acquisition

1 Introduction

The use of spoken language technology for language learning started in the late 1970s, but nowadays the so called CALL (Computer-Assisted Language Learning) systems are witnessing a great development due to the fact that speech technologies have made unprecedented progress and have achieved some kind of stability. CALL applications include a wide range of ICT applications, from the "traditional" drill-and-practice programs of the 1960s and 1970s to the recent applications in virtual learning environments and web-based distance learning. CALL also extends to the use of interactive whiteboards [1], computer-mediated communication (CMC) [2], language learning in virtual worlds, and mobile-assisted language learning (MALL) [3].

Naturalistic, implicit learning is not always enough to achieve high-quality L2 proficiency, according to second language (L2) acquisition research theories. Explicit instruction helps to overcome some of these learning problems [4][5]. This is the reason why nowadays software for CALL systems includes all types of material, mostly audiovisual. Nevertheless, the actual turning point in which CALL systems began to be more useful and sophisticated and came into more extended use was the implementation of speech technologies in them, especially automatic speech recogni-

tion (ASR). This fact allowed the interaction with users and the provision of automatic feedback, something absolutely essential in the process of learning a language. The modalities of interaction with the student comprise detection and assessment of pronunciation errors, perception training and use of talking heads, and detection and correction of prosody errors [6]. ASR systems are being used, above all, for pronunciation assessment; however, there are many other applications designed specifically over an ASR to provide intelligent feedback on important aspects of L2 learning such as morphology and syntax [7, 8], by means of using utterance verification techniques along with a predefined list of possible (correct and incorrect) responses for each exercise.

In this paper an ASR-based message verification system for Basque is presented. In this system the user's utterance is checked word by word in real time. The system displays a positively verified word in the same instant that it has just been uttered; otherwise, it waits until the expected correct word arrives. This system shows to be useful for exercises or tasks where the user is intended to choose or create an answer that has a strict word order, such as reordering sentences, answering questions and so on. Although the verification process can be carried out at word level or phoneme level, in this paper word level analysis is presented.

The paper is organized as follows: after the introduction, the theoretical basis in which the message verification system relies on is described. Then, several experiments and their corresponding results are presented. Finally, some conclusions and a reflection about development, improvement and future work are presented.

2 The basis of the system

2.1 The database

Currently there is no suitable database for the development of CALL systems in Basque. Regarding ASR databases, the only one that is publicly available for Basque is a SpeechDat database recorded over the fixed telephone network [14], at 8 kHz and 16 bits. As the recording conditions of this database are very different of the ones that can be expected in CALL systems, this database was not appropriate for our experiments. The database selected for the experiments in this paper, which at the moment is only available for research, is a Speecon-like one, recorded at 16 kHz and 16 bits using one headset and one desktop microphone. In the experiments presented in this paper, only the part recorded by means of the headset microphone has been used. The database contains recordings from 230 speakers, both native and non-native, as well as dialectal and standard Basque data for the formers. The native speakers' subcorpus is composed by 149 speakers, and the non-natives' subcorpus includes 81 speakers who speak Basque as L2 at different levels. All this information is labeled and can be easily extracted from the textual data files.

The audio files have associated their corresponding orthographic transcription file, and a rule-based P2G transcriptor for Basque has been used to obtain their standard Basque phonetic transcription. The HMMs were trained using only the subcorpus of

native speakers, leaving the non-native speakers' part for future research about improvements and adjustments for real implementations of CALL systems. Two thirds of the native speakers were used to train the acoustic models and the remaining third for testing. A mean of 170 files have been used per speaker. 60 of these files contain elicited speech, where the speaker is asked to read different types of texts (dates, numbers, phonetically rich sentences and the like), and the remaining 110 are commands, which are composed mainly by isolated words.

2.2 The ASR and calculation of PS scores

The message verification system has been built on the ASR-based utterance verification system developed at the Aholab Signal Processing Laboratory [9]. The system relies on a standard ASR based on Hidden Markov Models (HMMs). It processes 16 kHz signals with 16 bits and extracts vectors of 39 MFCC (Mel-frequency cepstral coefficients) —including first and second derivatives— from 25 ms duration frames each 10 ms. The decoding process is carried out by means of the Viterbi algorithm over an HMM lattice. The HMMs are context-dependent (triphones).

The verification process needs a second HMM lattice running in parallel, so that the *Goodness of Pronunciation* (GOP) scores of a phoneme y_u are computed as its posterior probability, over the acoustic segment X_u provided by the Viterbi decoder (u denotes the phoneme index). The parallel lattice consists of a free loop of context-independent HMMs, in order to avoid excessive increase of the processing time. Thus, equation 1 is used to calculate the GOP score of a phoneme:

$$GOP(y_u) = \log \Pr(y_u / X_u) \approx \frac{1}{T_u} \cdot \log \left[\frac{p(X_u / y_u) p(y_u)}{\sum_{k=1}^{N} p(X_u / y_k) p(y_k)} \right] \approx \frac{1}{T_u} \cdot \log \left[\frac{p(X_u / y_u)}{p(X_u / y_{j_{\max}})} \right] \quad (1)$$

where N is the number of phonemes and j_{max} is the phone model that gives the highest likelihood for the given segment. The denominator in equation 1 is replaced by the Viterbi likelihood of the segment given by the phoneme loop. Although some refinements are used to further improve the scores, many works show that this is a good confidence measure [10, 11].

The overall phoneme score (PS) for a word can be readily defined as a weighted sum of the normalized GOPs of its composing phonemes:

$$PS(word) = \sum_{k=1}^{N} w_k \cdot GOP(phoneme_k) \quad (2)$$

where w_k is the weight of the k-th phoneme among the N phonemes composing the word. Typically, the weights are equal for all the phonemes [12].

An example of the behavior of the PS scores can be seen in Fig. 1. The analyzed utterance contains three words separated by silences. The PS scores have been calculated just at the output state of the final HMM of each word for each frame. The three PS score sequences are represented with different line types, over the spectral repre-

sentation of the signal, in order to have a better idea of the limits of each word. As can be seen in the figure, each line reaches its maximum (of the –PS curve) when the corresponding uttered word finishes. That leads us to think that the three words would be correctly detected when the –PS score of each of them reaches a maximum over a certain threshold (around zero in the figure). A similar analysis could be done at phoneme level, where the PS score sequence would be calculated at the output state of each HMM for each frame.

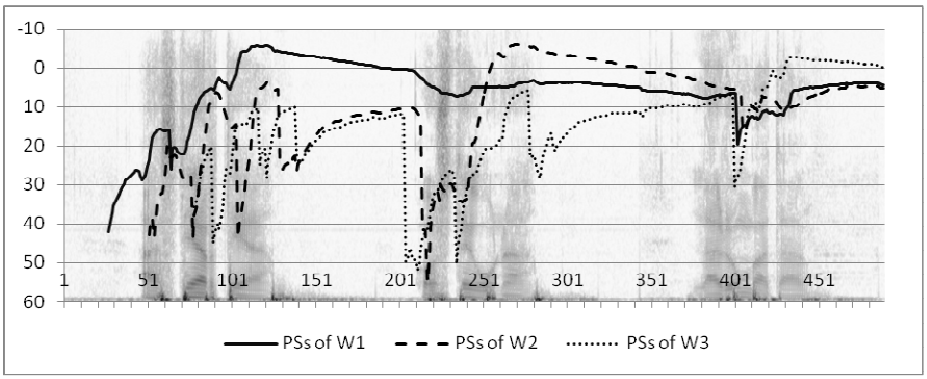


Fig. 1. PS score sequences of three words: W1 ("asteartea", *Tuesday*), W2 ("osteguna", *Thursday*), W3 ("larunbata", *Saturday*), over the spectral representation of the signal.

2.3 The decision thresholds

The main problem that arises when working with both GOP and PS scores is how to calculate the decision thresholds. Two distributions are needed for this purpose: on the one hand, the distribution of the values of the PS scores when the expected or correct utterance is being verified and, on the other hand, the distribution of the values of the PS scores when an unexpected or incorrect utterance (for instance a different word or part of a word) is being verified. Then, a cut point between the two distributions can be selected, e.g. the equal error rate (EER), at which both the probability of false acceptance and false rejection are equal. This threshold may be moved taking into account the level of the L2 student; for a beginner, for example, a larger amount of errors can be accepted and, on the contrary, the system should be stricter with an experienced and high-skilled student.

The PS scores of the correct utterances can be calculated using the segmentation provided by the ASR in forced alignment mode. In order to calculate the incorrect utterances' PSs, one valid technique is to artificially introduce errors in the dictionary. This is particularly useful if we consider that, in general, scarcity of data is a common problem in this kind of research [9, 13]. The literature shows that this technique has been successfully used to calculate phoneme GOP distributions. In this paper word level scores will be calculated and assessed, since it can be useful for tasks where not so strict results are needed, as working with beginners. The way to introduce errors in

the dictionary consists in substituting a word by any other word in the sentence grammar, maintaining the segmentation obtained by a previous evaluation of the correct word sequence. Thus, the utterance that the system receives does not fit with the one that it is expecting, and so we can consider that for the system the audio file corresponds to an incorrect utterance.

The histograms of both the PS distribution obtained using this procedure for incorrect utterances and the PS distribution of the correct utterances are shown in Fig. 2. PS computation for incorrect utterances has been carried out three times, in order to obtain sufficient data, since the amount of incorrect words is one per file. The EER is located in the point 0.375, with a value of 2.12 %. This value is a priori encouraging, since the error made classifying a new incoming PS score can be considered very small.

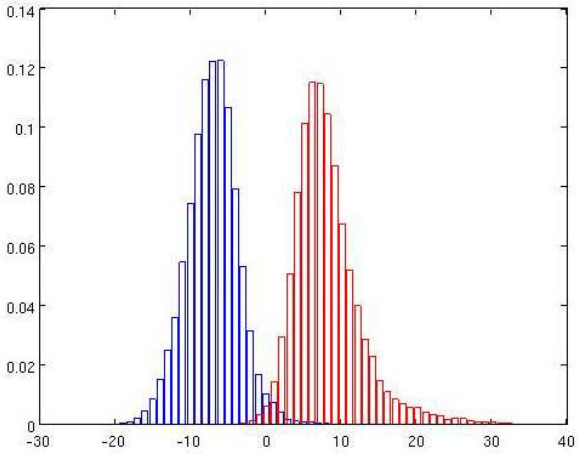


Fig. 2. The normalized histograms of the PS distributions of correct (left) and incorrect (right) utterances.

3 Experiments and results

The measure selected to assess the results is the widely used *SA* coefficient (Scoring Accuracy), which is calculated as in equation 3.

$$SA = \left(\frac{(CA + CR)}{(CA + CR + FA + FR)} \right) * 100 \quad (3)$$

where: *CA*: Correctly Accepted; *CR*: Correctly Rejected; *FA*: Falsely Accepted; *FR*: Falsely Rejected.

Two different experiments have been carried out in order to check the consistence of the calculated decision threshold. The first experiment consists in observing whether the utterances that the system is expecting are labeled as correct (*CA*) or as

incorrect (*FR*). For this purpose, the part of the database left for testing has been used: 2,218 files in total, with an amount of 7,296 words. The decoding was performed by the ASR in forced alignment mode, where two options were taken into account between words: an optional silence or a coarticulated transition. The PS scores were obtained over the final segmentation of the Viterbi decoder.

The results obtained are shown in the first result row of Table 1. We can see that SA is 97.18 % or, considering the error instead of the accuracy, 2.82 % error rate is achieved, which coincides approximately with the value of EER obtained for the decision threshold. It is worth mentioning that 75.24 % of the FR words contain 3 phonemes or less. So, it is evident that, as one could expect, the longer the word is, the more robust the result is. This may be due to the fact that if one of the phonemes is not correctly pronounced it affects more to the overall scoring in a short word than in a longer one. As an example, we noticed that in the results there are many short words containing the character *j*, which in standard Basque must be pronounced as /*jj*/ but in many dialects is pronounced as /*x*/[1].

Table 1. Results of the experiments 1 and 2

	CA	CR	FA	FR	SA
Experiment 1	7,090	---	---	206	97.18 %
Experiment 2	---	1.174	0	---	100.00 %

For the second experiment, only isolated words have been taken into account. In this case, artificially inserted errors consist in randomly substituting the input textual word with another word in the dictionary. The files containing isolated words in the test part of the database are 1,174, and the PS score provided over the ASR segmentation shows that none of them is classified as correct. So, a scoring accuracy of 100 % is obtained, as can be seen in the second result row of Table 1.

4 System design

In a realistic environment the student will make mistakes. That means that the verification system must be able to manage these extra voice segments in order to absorb the effects on the Viterbi decoder. Since the system will receive more voice frames than it expects, this must be modeled somehow. So, in the design of the final system an optional phoneme loop has been added to the decoding lattice of the ASR at the beginning, at the end and between words, in the way shown in Fig. 3. If this phoneme loop was not added, the segmentation resulting from the Viterbi algorithm would not be predictable, and the verification or scoring could not be calculated over this segmentation.

The system will be first assessing the PS scores for the first word W1, frame by frame (each 10 ms), until one of them goes over the threshold calculated previously as indicated in section 2.3. The PSs will be calculated over the segmentation that the

[1] http://aholab.ehu.es/sampa_basque.htm

incoming Viterbi token has in the last HMM state of the word. Then, the system will wait for a maximum (in the –PS curve), checking that the scores of the next *N* frames are smaller than this value. When a maximum is detected in that way, the word will be displayed for the user, and the same process will start for the next word W2.

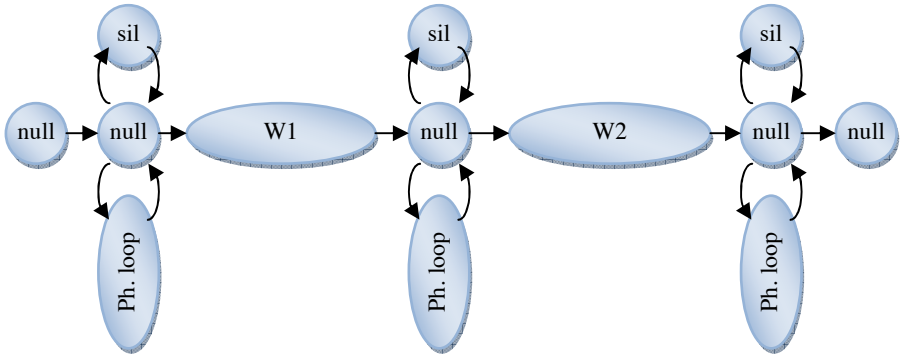


Fig. 3. The decoding lattice scheme for a sentence of two words, with optional silences and phone loops between words

In order to assess the designed system in a more realistic environment, a new experiment has been carried out, where expected and unexpected speech are combined. The test has been devised as follows: for each sentence or word sequence to be evaluated, one word has been deleted in the transcription. The aim of this experiment is to simulate a situation where a user utters some words correctly, in the same order as expected, then an incorrect word, and finally the remaining words correctly again. Notice that the word that is being evaluated during the incorrect segment is the next word that the system expects. So, we can obtain the score of that word when an incorrect word is uttered, and afterwards the score of the same word when the correctly uttered word arrives.

886 sentences have been used in this experiment, containing 5,080 uttered words in total. Each sentence ranges from two to seventeen words. In each written input sentence one word has been deleted, so that in the gaps created by the deletion the next word is verified. Thus, the word following to a deletion is verified twice: at first while the deleted word is being pronounced, and then when its corresponding utterance arrives. Thus, the PSs of 4,194 words —which are considered as correct— will be evaluated, and, in addition, 886 of them will be also evaluated for the unexpected speech segment, 5,080 scores in total.

The SA obtained in this experiment is, as can be seen in Table 2, 96.63 %. Regarding just the unexpected or incorrect words, a scoring accuracy of 84.88 % has been obtained (752 out of 886). Nevertheless, the total amount of incorrect words (886) is smaller than that of the correct ones (4,191) and that is why their influence is not so evident on the global rate.

Table 2. Results of the experiment simulating errors

Experiment results	
CA	4,157
CR	752
FA	134
FR	37
SA	**96,63 %**

5 Conclusions and future work

In this paper the design of a message verification system is introduced suitable to be implemented in CALL systems, in tasks that need word-by-word verification in real time. As soon as a word is detected, it is displayed to the user. In such a system, rejecting unexpected speech is as necessary as detecting correct words (words that the system is expecting). In this paper we describe a way to calculate the decision threshold, inserting artificial controlled errors. The results of the experiments show that the system has better scores when accepting expected words than when rejecting unexpected speech, although the decision threshold has been calculated using the EER. So, further analysis must be done, in order to detect why this asymmetry happens and adjust the decision threshold.

An experimental approximation to a realistic environment has been applied to better evaluate the system, but it lacks of real users' evaluations. Thus, some more tests must be devised to complete the analysis and to assess the system in a real environment. It would be interesting to test the system with students in different points of their language acquisition process, in order to adjust the thresholds to different situations. Another interesting experiment to do is to perform phoneme-level verification and compare the results of both systems.

The system presented in this paper has been designed for Basque, since Basque acoustic models have been used. However, the strategy can be easily followed to develop a message verification system for any other language, by means of creating the corresponding acoustic models for that language.

6 Acknowledgements

This work has been partially supported by the Basque Government (Berbatek, IE09-262), the University of the Basque Country (UE-08/11) and the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02).

References

1. Schmid, E.C.: Interactive whiteboard technology in the language classroom: exploring new pedagogical opportunities, Saarbrücken, Germany: VDM Verlag Dr. Müller (2009)
2. Lamy, M.N., Hampel R.: Online communication in language learning and teaching, Houndmills: Palgrave Macmillan (2007)
3. Shield, L., Kukulska-Hulme, A.: Special edition of ReCALL (20, 3) on Mobile Assisted Language Learning (2008)
4. Norris, J.M., Ortega, L.: Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis, Language Learning, vol. 50, pp. 417-528 (2000).
5. Ellis, N.C., Bogart, P.S.H.: Speech and Language Technology in Education: the perspective from SLA research and practice, In Proceedings ISCA ITRW SLaTE, Farmington PA (2007)
6. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10): 832–844 (2009)
7. DISCO project website, http://lands.let.ru.nl/ strik/research/DISCO/.
8. Van Doremalen, J., Strik, H., Cucchiarini, C.: Utterance Verification in Language Learning Applications. Proceedings of the SLaTE-2009 workshop, Warwickshire, England (2009)
9. Odriozola, I., Navas, E., Hernáez, I., Sainz, I., Saratxaga, I., Sánchez, J., Erro, D.: Using an ASR database to design a pronunciation evaluation system in Basque. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4122-4126 (2012)
10. Witt, S., & Young, S. J.: Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30(2-3), 95-108 (2000)
11. Franco, H.: Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30(2-3), 121-130 (2000)
12. Mak B., Siu M., Ng M., Tam Y., Chany Y., Chan K., Leung K., Ho S., Chong F., Wong J., Lo J.: PLASER: Pronunciation Learning via Automatic Speech Recognition. In Proc. of HLT-NAACL, May, 2003, Edmonton, Canada (2003)
13. Kanters, S., Cucchiarini, C. and Strik, H.: The Goodness of Pronunciation algorithm: a detailed performance study, In Proceedings of SLaTE 2009, Birmingham (2009)
14. Hernaez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., Sanchez, J.: The Basque speech_dat (II) database: a description and first test recognition results, In Eurospeech-2003, 1549-1552 (2003)

Vocalic Filler Characterization using Acoustical Properties Derived from Cepstrum

Arlindo Veiga[1,2], Jorge Proença[2], Dirce Celorico[2], Sara Candeias[2], Fernando Perdigão[1,2]

1 Electrical and Computer Eng. Department, University of Coimbra, Portugal
2 Instituto de Telecomunicações - Polo de Coimbra, Coimbra, Portugal

{aveiga,jproenca,dircelorico,saracandeias,fp}@co.it.pt

Abstract. The study presented here attempts to characterize the so-called vocalic fillers, according to their acoustic and prosodic properties. A comparison with the vowels of the phonemic system of the European Portuguese possessing similar timbre is also provided. Among the acoustic and prosodic parameters used to characterize such speech segments, in this contribution we paid particular attention to the spectral shape (timbre) and duration. Such parameter analysis was conducted by means of training HMM models, as is usually done for speech recognition purposes, from which a spectrum is derived. This is an extension of a previous work towards the characterization of the hesitation phenomena with a new methodology. The corpus used is a subset of the Portuguese television broadcast news (RTP) and of the Tecnovoz speech database. This study allowed us to characterize two of the most common vocalic fillers in Portuguese spontaneous speech as well as the two unstressed vowels with similar timbre.

Keywords: vocalic fillers, unstressed vowels, formant extraction, spontaneous speech

1 Introduction

The quick development of speech systems has led to the challenge of dealing with a variety of speech events occurring in spontaneous speech. In order to increase the development of successful and useful speech technology, the description and characterization of the spontaneous speech must be addressed. In this context, several works on hesitation phenomena have gained importance over the last years. Furthermore, studies that provide valuable information on hesitation phenomena made it possible to understand the speech structure and the speaker's communicative intention [1]. Filled pauses or vowel lengthening, as well as repetitions, restarts and repairs have been described as hesitations in studies such as [2-4]. For what concerns the studies on Portuguese, contributions such as [5-8] have already used some acoustic characteristics to classify hesitation events in spontaneous speech. However, and although some recent research on hesitation phenomena for Portuguese can be found (for instance,

[7]), most of the studies have been focused mainly in English [2], Swedish [3] or French [4].

In our previous work for European Portuguese (EP) [5], we analyzed the regular patterns of filled pauses and extensions. In the present study, we aim at extending our research in a more demonstrative and quantifiable way. We are now focusing mainly on the phenomenon of fillers corresponding to autonomous vocalic fillers [8]. Such events have no lexical support and are linguistically distinct from the vowels belonging to a lexical segment (mainly to the vowels <a> pronounced as [ɐ] or <e> pronounced as [ə] (schwa), either they are placed in an intra-word position or are functional words themselves, such as the definite articles or the conjunctions).

Specifically, the goal of our contribution here is to acoustically and prosodically characterize the most common vocalic fillers in EP [5] (corresponding to [ɐ] and [ə]), by using a different approach of analysis. Based on the acoustic typology for investigating the vocalic particularities of the autonomous fillers in a multilingual perspective [8-9], we focus on the timbre (F1/F2) and duration of the vocalic fillers. Such parameter analysis has been commonly conducted using Praat [10]. Here we used a different methodology by converting MFCC (Mel-Frequency Cepstral Coefficients) cepstral information into spectrum. We also evaluate these parameters in the framework of the lexical vowels in the EP, in order to compare the vocalic segment of the hesitations with the lexical segments having similar timbre.

This paper is outlined as follows: Section 2 describes the corpus and the adopted methodology; Section 3 is dedicated to the reconstruction of the spectrum from MFCC coefficients and is followed by a discussion of the results (Section 4). Finally, Section 5 summarizes the current findings.

2 Corpus and methodology

Tecnovoz speech corpus and audio Broadcast News corpus (BN-RTP) were used as a source of the speech segments needed for this study. The two corpora have different speech styles. Speech in Tecnovoz contains reading prompted text without fillers (see [11] for a more detailed description). The BN-RTP corpus corresponds to January-February 2011 TV programs, totalizing about 22 hours, from which vocalic fillers were manually transcribed. The speech segments analyzed in the present study are [ɐ] and [ə]. Table 1 shows the number of the speech segments that occur in the two corpora. Tecnovoz corpus was used previously for training a speech recognition system, justifying the higher number of segments in this corpus. Recognition results are presented in [11] and [12]. Previous results on filled pauses as well as on vocalic extensions can be found in [5].

Vocalic fillers and lexical vowels differ considerably, for instance on the average of the length and prosodic properties. In order to characterize such segments, we can use the vowel models previously constructed and follow a similar approach for the vocalic fillers. We have adopted this method, which allows direct comparisons, and we have trained acoustic models for the vocalic fillers with the same parameterization. The parameters are the MFCC which are commonly used for speech recognition

purposes. Only one state is used in these models due to the constancy of these segments.

The cepstral representation has some advantages, at least for speech recognition, allowing a smooth spectral representation, modeling speaker and environment variability and retaining perceptual relevant information for recognition. This is a valid alternative of the formant analysis commonly conducted using Praat [10]. The disadvantage of this method is that we need an algorithm to convert MFCC cepstral information into spectrum. The next section deals with this problem.

Table 1. Information on the Tecnovoz and BN-RTP corpora, including the number of occurrences of the speech segments considered.

Vowel/Filler	Tecnovoz	BN-RTP
[ɐ]	149372	246
[ə]	41408	112

3 Spectrum reconstruction from MFCC coefficients

The problem of obtaining a spectral estimate from the Mel-Frequency Cepstral Coefficients has been addressed several times in the past (see for instance, [13-15]). In the present case we intend to describe phone characteristics through the use of the acoustic models obtained when training a speech recognition system. These acoustic models are based on Mel-Frequency Cepstral Coefficients obtained with a traditional front-end: pre-emphasized speech, windows of 25ms every 10ms, about 30 triangular filters, no zero order coefficient c0 (which is replaced by the frame log energy), 12 MFCCs, and, many times, cepstral mean normalization. The GMM (Gaussian Mixture Model) of the middle state of a phone model is taken in order to describe its acoustic characteristics. The cepstral information is transformed into a corresponding smooth spectrum using only the mean vectors and the mixture weights.

Mostly, we follow the method described in [13] and [15], although with some variations which have not been published yet, as far as we know. The method is the following. From the ceptral vector $\mathbf{c}=[c_1,c_2,\ldots,c_{12}]^T$ we obtain a smooth version of the log energies of the outputs of the mel-filter bank, corresponding to the truncation of the cepstrum:

$$\hat{Y}[k] = \sum_{n=1}^{12} \sqrt{\tfrac{2}{N}} \cos\left(\frac{\pi}{N} n(k-\tfrac{1}{2}) \right) \times c[n] = \mathbf{D}_1^T \cdot \mathbf{c} \quad , \quad k \quad 1..N \, , \qquad = \qquad (1)$$

where $\mathbf{D}_1$ is the DCT sub-matrix obtained by removing the first and the 13th to Nth lines, and where N is the number of filters in the triangular filter-bank (28 in our case). We intend to obtain an estimate of the spectrum in $\frac{N_{FFT}}{2}+1$ linearly spaced points in order to do DFT operations. Instead of using an expansion factor (as in [15]), we can make the previous equation continuous in k, so that for an integer k the result

is the same as before. As the filter apex frequencies are equally spaced in a mel scale and they are taken in mels, as

$$mel\left(f_c[k]\right) = \frac{mel(f_s/2)}{N+1}k,\ \ k = 1..N\,, \tag{2}$$

where f_s is the sampling frequency, we can use this equation for any frequency between 0 and $f_s/2$ in order to choose the appropriate k, possibly not integer, to use in the interpolation. For a DFT operation we wish to choose the frequency points as $f_k = k\frac{f_s}{N_{FFT}}, k = 0..\frac{N_{FFT}}{2}$, so that we can use a value of λ_k, to replace k, as

$$\lambda_k = (N+1)\frac{mel\left(f_k\right)}{mel(f_s/2)} = (N+1)\frac{\ln\left(1+\frac{f_k}{700}\right)}{\ln\left(1+\frac{f_s/2}{700}\right)}\,. \tag{3}$$

The required interpolated values of the linear spectrum can then be obtained using the equation

$$\tilde{Y}[k] = \sum_{n=1}^{12}\sqrt{\tfrac{2}{N}}\cos\left(\frac{\pi}{N}n(\lambda_k - \tfrac{1}{2})\right)\times c[n]\,,\ \ k \quad 0..\frac{N_{FFT}}{2}\,.\ \ = \tag{4}$$

Only a DCT sub matrix of dimension $\left(\frac{N_{FFT}}{2}+1\right)\times 12$ is needed to do the interpolation.

3.1 The effect of the missing c_0

We notice that the c0 value in this approach is implicitly zero, because $\sum_{k=1}^{N}\hat{Y}[k] = 0$, which can be easily verified using the orthogonality of the DCT matrix. However, the value of c_0 can be recovered from the log energy of the frame, $\log(E)$, usually used as a replacement of c_0. From the DCT equation, we have

$$c_0 = \frac{1}{\sqrt{N}}\sum_{k=1}^{12}Y[k] \tag{5}$$

where $Y[k]$ is the log output energies of the filters. We assume that the log energy value that is appended to the cepstral vector is equal to the sum of the output energies of the filters, specifically, $E = \sum_{k=1}^{N}\exp\left(Y[k]\right)$. In this case, computing

$$\hat{E} = \sum_{k=1}^{N}\exp\left(\hat{Y}[k]\right), \tag{6}$$

the log ratio, $\log\left(E/\hat{E}\right)$, corresponds to the missing log energy value to add to $\hat{Y}[k]$ in order to obtain the scaling present in $Y[k]$. The c_0 value can then be recovered as

$$\hat{c}_0 = \sqrt{N}\left(\log(E) - \log(\hat{E})\right). \tag{7}$$

Tests made with c0 and this recovered version of it, show that they are almost indistinguishable.

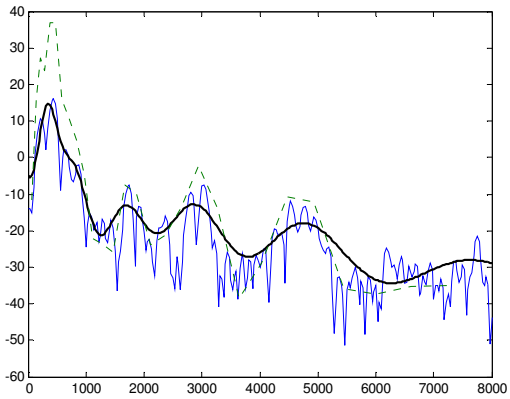


Fig. 1. Log spectrum of a segment of a vowel [ə] uttered by a female speaker (thin line), the corresponding log energies, *Y*[*k*], for the 28 filters (dotted line) and the interpolated spectrum with 257 points obtained from the MFCC cepstrum (thick line).

3.2 The effect of pre-emphasis and filter-bank

There are also two other factors to consider for the interpolated spectrum $\tilde{Y}[k]$ to be a faithful version of the original spectrum: the effect of the filter-bank filters and the pre-emphasis applied to the original signal. Since this corresponds to a cascade filter, the corresponding cepstrum is a sum of the original cepstrum with the cepstrum of the pre-emphasis filter, as noticed in [13]. Assuming an impulsive input (a constant input spectrum) these two effects can be taken at once, computing this pre-emphasis cepstrum with unitary input spectrum, $c_p[n]$, that must be subtracted to the original cepstrum $c[n]$. This must be done even if the pre-emphasis factor is zero.

If a cepstral mean normalization was used in the analysis front-end, corresponding to the subtraction of the mean cepstrum, then we need this mean cepstrum in order to recover the original spectrum. In general, this does not have a very dramatic effect unless the mean log spectrum has a large variance.

Figure 1 shows an example of a DFT log spectrum in dB of a segment of a phone [ə] uttered by a female speaker, as well the 28 values Y[k] and the interpolated spectrum, $\tilde{Y}[k]$, with 257 points (thick line).

3.3 The spectrum pattern of the phones and fillers

The spectral characterization of a phone or filler is embedded in the GMM of their states. We take the middle state as the characterization of the stable portion of the phone. The various components of the mixture correspond to different realizations of the phone by different speakers, acoustic conditions and coarticulation conditions. Figure 2 shows the log spectrum of the 10 highest weighted mixture components for the phone [ə], obtained with the previously explained method. The thick line represents the average weighted log spectrum, which corresponds to the sum of each mean component multiplied by its mixture weight. Figure 3 shows the same for the phone [ɐ].

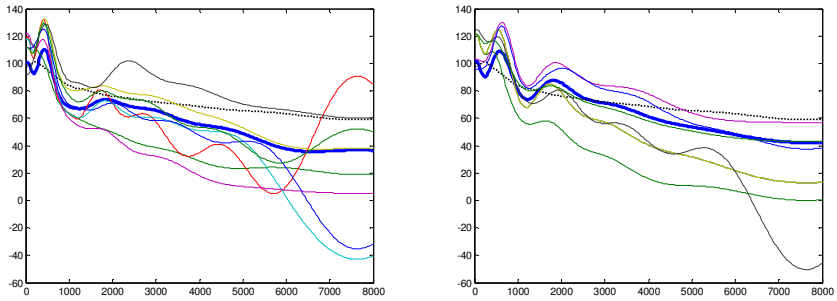
4 Results and discussion

Figures 2 and 3 show the spectral information patterns from which the formants can be derived. Figure 4 shows the normalized histogram of the duration of the segments. With this set of features, we aim at characterizing vocalic fillers. Results partially agree with the ones reported in [16], mainly regarding the formant values of the schwa [ə]. However, our results can be considered relevant, due to the larger number of segments analyzed.

As can be seen in Figure 2a), the first and second formants are evident in all mixture components. F2 varies more than F1 as a result of different speakers. In comparison with Figure 3a) it is shown that the main spectral characteristics of the filled pause and the schwa are similar in most components. However, as it can also be seen, there are mixture components of the vowel [ə] that closer resemble a fricative than a vowel. In reality, this corresponds to the fact that in most cases the [ə] is pronounced without any vibration of the vocal folds. This is indeed characteristic of the schwa (see [18] for instance). Obviously, this does not happen in the filled pause realized with the timbre of [ə]. While the first formant is quite stable in both vocalic filler and lexical vowel, the second formant varies more when it corresponds to a lexical vowel, reflecting the fact that hundreds of speakers (about 300) are present in the database. This contrasts with many phonetic studies, where only one or few speakers are considered.

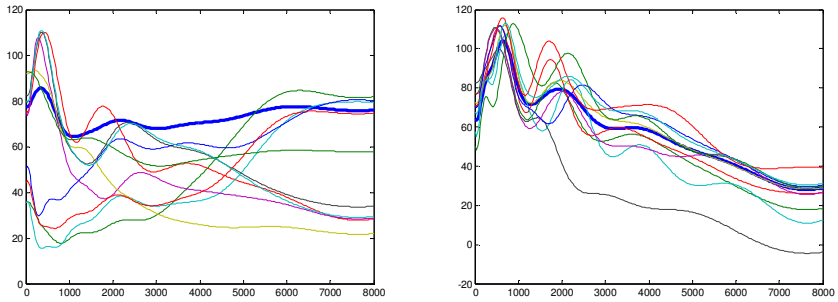
Figure 2b) and 3b) show the results for the vocalic filler and lexical vowel [ɐ]. Concerning [ɐ], the similarity between filler and the vowel is even more evident than in the [ə] case.

As presented in [5], F0 is almost constant during filler segments, which is also a strong indicator of the presence of a hesitation phenomenon. As expected, the duration of the fillers is much higher than in the case of the lexical related vowels. This is indicated in Figure 4.



a) b)

Fig. 2. a) - Log spectrum of the vocalic filler, with timbre [ə], from a GMM with 8 mixture components. Dotted line: spectrum derived from the mean cepstrum. F1: 400 - 470 (mean: 430) Hz; F2: 1700 - 2300 (mean: 1800) Hz. **b)** Log spectrum of the vocalic filler with timbre [ɐ]. F1: 400 - 630 (mean: 590) Hz; F2: 1700 - 2100 (mean: 1850) Hz.



a) b)

Fig. 3. a) Log spectrum of the vowel [ə] in Tecnovoz database, obtained from the second emitting state of the corresponding HMM model. The thin lines correspond to the 10 major mixture components and the thick line to the log spectrum from the weighted mean cepstrum (mean vector of the model). F1: 300 - 450 (mean: 370) Hz; F2: 1500 - 2500 (mean: 2100) Hz. **b)** Same as a) for the vowel [ɐ]. F1: 470 - 870 (mean: 660) Hz; F2: 1700 - 2200 (mean: 1900) Hz.

5 Conclusion

This study addresses the problem of characterizing the vocalic fillers in Portuguese spontaneous speech. The analysis is conducted both over Portuguese read-speech corpus (Tecnovoz) and BN-RTP corpus. Vocalic fillers and lexical vowels of the phonological system of European Portuguese possessing similar timbre are compared in terms of their acoustic and prosodic features. We aimed to extend our previous work on the characterization of hesitation phenomena. The acoustic characterization of the segments is based on their spectral properties, derived from models trained for

speech recognition. Thus, from the models we can successfully reproduce average spectral behavior of the speech segments. The duration is the characteristic that better differentiates vocalic fillers from the related lexical vowels.

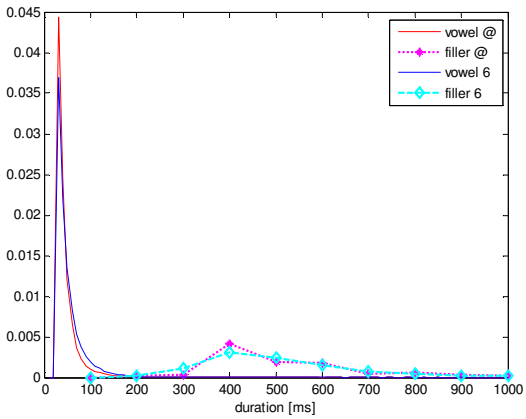


Fig. 4. Normalized histograms of the duration of the lexical vowels and the fillers with the same timbre (legend in SAMPA, [17]: [@]= [ə]; [6]= [ɐ]).

Extending this study on vocalic segments to other hesitation events, besides vocalic fillers, will allow for a better modeling of spontaneous speech. Insights into the process of human speech communication for Portuguese will be gained as well.

Acknowledgments. This work is funded by FCT and QREN projects (PTDC/CLE-LIN/11 2411/2009; TICE-Healty13842) and partially supported by FCT (Instituto de Telecomunicações multiannual funding PEst-OE/EEI/LA0008/2011). Sara Candeias is supported by the FCT grant SFRH/BPD/36584/2007. The authors would also like to thank Rádio e Televisão de Portugal (RTP) for his support with the multimedia corpora.

References

1. Levelt, W.: Speaking: From Intention to Articulation. MIT Press, Cambridge, Massachusetts (1989)
2. Shriberg, E.: Preliminaries to a Theory of Speech Disfluencies. Ph.D. thesis, Berkeley's University of California (1994)
3. Eklund, R.: Disfluency in Swedish human–human and human–machine travel booking dialogues. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Department of Computer and Information Science, Linköping University, Sweden (2004)
4. Candea, M.: Contribution à l'étude des pauses silencieuses et phénomènes dits "d'hésitation" français oral spontané. PhD thesis, University of Paris 3-Sorbonne Nouvelle (2000)

5. Veiga, A., Candeias, S., Lopes, C., Perdigão, F.: Characterization of hesitations using acoustic models. In: Proc. of the 17th International Congress of Phonetic Sciences (ICPhS XVII), 2054-2057, Hong Kong (2011)
6. Veiga, A., Candeias, S., Celorico, D., Proença, J., Perdigão, F.:Towards Automatic Classification of Speech Styles. In: the 10th International Conference on Computational Processing of Portuguese (PROPOR 2012), Coimbra, Portugal (2012). In: H. Caseli et al. (eds.). Lecture Notes in Artificial Intelligence (LNAI) 7243: 421-426, Springer-Verlag Berlin Heidelberg (2012)
7. Moniz, H., Trancoso, I.,Mata, A.: Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In: Proc. of Interspeech'09, 1719-1722, Brighton (2009)
8. Candea, M.,Vasilescu, I., Adda-Decker,M.: Inter- and intra-language acoustic analysis of autonomous fillers. In: Proceedings of Disfluency In Spontaneous Speech (DISS) Workshop, Aix-en-Provence (2005)
9. Vasilescu, I., Adda-Decker,M.: Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech. In: INTERSPEECH 2006, Pitsburgh (2006)
10. Boersma,P.,Weenink,D.: Praat, a system for doing phonetics by computer. Institute of Phonetic Sciences of the University of Amsterdam, pp. 132-182 (1999)
11. Lopes, J., Neves,C., Veiga, A., Maciel, A. M., Lopes, C., Sá, L. V., Perdigão, F.: Training A Robust Command Recognizer with the Tecnovoz Database. In: Proc Jornadas en Tecnología del Habla - JTH, Bilbao, Spain (2008)
12. Veiga, A., Candeias, S., Perdigão, F., Sá, L.: Using Coarticulation Rules in Automatic Phonetic Transcription. In: 9th International Conference on Computational Processing of Portuguese (PROPOR 2010), April, Porto Alegre - RS, Brazil (2010)
13. Milner, B. P., Shao, X.: Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. In: 7th International Conference on Spoken Language Processing (ICSLP-2002), September 16-20, Denver, Colorado, USA (2002)
14. Boucheron,L. E., De Leon, P. L.: On the Inversion of Mel-Frequency Cepstral Coefficients for Speech Enhancement Applications. In: Proc. Int. Conf. Signals and Electronic Systems (ICSES), pp.485-488 (2008)
15. Ramabadran,T., Meunier, J., Jasiuk, M., Kushner, B.: Enhancing distributed speech recognition with back-end speech reconstruction. In: Proc. Eurospeech, pp. 1859-1862 (2001)
16. Andrade, A.: Reflexões sobre o 'e mudo' em Português europeu. In: I. Duarte and I. Leiria (eds), Proceedings of the I International Congress on Portuguese, Vol III, Abril 1994, U. Lisbon, Colibri/APL, pp.303-344,Lisbon (1996)
17. Wells, J.C.: SAMPA Computer Readable Phonetic Alphabet. In: Gibbon, D., Moore, R. and Winski, R. (eds.), Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV (1997)
18. Candeias, S., Veiga, A., Lopes, C., Perdigão, F.: A realização do schwa no Português Europeu. In: II workshop on Portuguese description-JDP (Satellite Workshop of STIL 2011), October 24-26, Cuiabá, UFMG – Brazil (2011)

Special Session Projects

Architecture for Text Sign Localization and Recognition

Arturo Romero, José Moreno, Fernando Fernández, Juan Manuel Montero, Rubén San-Segundo

Grupo de Tecnología del Habla. E.T.S.I.Telecomunicación.
Universidad Politécnica de Madrid

arturoromero@die.upm.es

Abstract. This paper describes a low complexity strategy for detecting and recognizing text signs automatically. Traditional approaches use large image algorithms for detecting the text sign, followed by the application of an Optical Character Recognition (OCR) algorithm in the previously identified areas. This paper proposes a new architecture that applies the OCR to a whole lightly treated image and then carries out the text detection process on the OCR output. The strategy presented in this paper significantly reduces the processing time required for text localization in an image, while guaranteeing a high recognition rate. This strategy will facilitate the incorporation of video processing–based applications into the automatic detection of text sign similar to that of a smartphone. These applications will increase the autonomy of visually impaired.

1 Introduction

The partial or complete loss of vision is a challenge that some people are forced to confront during their lives. Representing visual information using alternative senses (such as hearing or touch) may be a useful strategy in order to make easier some situations that visually impaired people have to deal with in the daily life. Detecting visual information automatically (using visual signal processing technologies) involves significant challenges in order to be useful for blind people. One of the most outstanding ones is the high power consumption for processing the visual information in real time. Because of this, many systems developed so far focus on specific detection problems such as the reading of car license plates, the shape of some objects, etc. One of the most time consuming processes is to locate, in the visual image, the target object (i.e. car license plate) before extracting its shape or content.

This paper presents the design and implementation of a system for detecting and recognizing text signs automatically and translating them into speech. This system focuses on relevant information that visually impaired people cannot read by themselves, for instance, the emergency exit sign in a building or the direction signs in the street. For these kinds of applications, the system has to be included in a mobile device such as a smart-phone. Because of these constraints, it is necessary to implement reduced detection algorithms in order to make the system work in real time. Instead of

using lengthy image algorithms for detecting the text sign and then applying an Optical Character Recognition (OCR) algorithm in the previously identified areas, this paper proposes the application of the OCR in a whole lightly-treated image followed by the text detection process on the OCR output. The OCR output is a text output, so its processing is less time consuming.

When applying the OCR on a whole image, the OCR focuses on the features of the text itself, such as corners, strokes, edges, etc. Because of this, the OCR introduces an elevated number of mistakes giving rise to confusing elements in the image, for example the corner of a window, with a written character. The system proposed in this paper uses a localization algorithm, based on the idea of the text alignment methods that discharge the characters mistakenly introduced by the OCR, and detects the content of a specific text sign. Finally, the system will produce both an audible and an enhanced visual interpretation (for partially visually impaired people) of the sign.

This paper is organized as follows. Section 2 presents a state of the art on what has already been achieved in the area of text detection in images. The system proposed in this paper is described in Section 3. Finally, the implications and the main conclusions of the work are described in Section 4.

2 Previous work

Many text detection and localization approaches use features related to text characters (Chen et al, 2004). Then, the edge regions are grouped together based on the common features (Jirattitichareon, 2006).

Texture is another widely used feature for text detection and localization (Kim et al, 2003; Jung et al, 2000). First a texture analysis method such as Gabor filtering is used to extract the features of the texture. Then, a classifier is used to identify a region as a text or non-text based on these features. Support vector machines are used in (Kim et al, 2003; Jung et al, 2000) for this purpose. In (Huang and Ma, 2010), the authors use LoG-Gabor filters to obtain the stroke map and apply Harris corner detection to find the seed points for the connected component analysis.

Some approaches assume that the text is written in the horizontal or vertical direction (Sin et al, 2002; Wu et al, 2005). However, this constraint is not practical due to the possible tilt of the camera. Therefore, an affine rectification step is usually added to improve the results (Chen et al, 2004).

Text that appears in video frames is an important cue for the content analysis and indexing of the video. Accordingly, several approaches have been proposed for the detection and tracking of text in videos (Shivakumara et al, 2011; Yu and Wang, 2010). Some of these methods are used to detect the artificial text added to the video, based on the fact that this type of text has a homogeneous color as opposed to the complex color distribution of the scene (Yu and Wang, 2010). Other methods try to detect any type of text.

These methods usually have some common procedures for image treatment or text detection. Due to this situation, there are some implemented libraries or programs performing the basis image processing and character recognition.

A very well-known library for image processing is OpenCV (Open Source Computer Vision), provides a large amount of functions that allow the developers to work on a pixel level and implement higher level functions easily.

Among the character recognition world, Tesseract can be found as one of the best OCR programs. This open source engine was one of the top 3 engines in the 1995 UNLV Accuracy test. Between 1995 and 2006 it had little work done on it, but since then it has been improved extensively by Google and is probably one of the most accurate open source OCR engines available. Combined with the Leptonica Image Processing Library it can read a wide variety of image formats and convert them to text in over 40 languages (http://code.google.com/p/tesseract-ocr/).

3 Proposed approach

The system proposed in this paper can be divided into two big sections: hardware and software. The hardware section consists of a portable device and a camera, preferably being integrated at the back of the chosen device in order to locate text signs in one movement, if there are any. The software section is the main one in this project. A schematic representation of the process used to detect text signs are shown in Fig 1.

As shown in Figure 1, the system captures the image from the camera and performs an image treatment for converting a color image into a Black & White (B&W) image. Afterwards, the OCR algorithm looks for the text patterns contained in the whole B&W image, saving the text output into a .txt file. The text localization algorithm then detects whether or not the file given by the OCR contains any of the patterns included in the text sign database. In a text sign database, the system stores all possible text signs that can be detected and recognized by the system. If any of the text patterns stored in the database is found, the system will provide both an audible and an enhanced visual representation. If not, the process for sign searching starts again with a new frame captured after a set time.

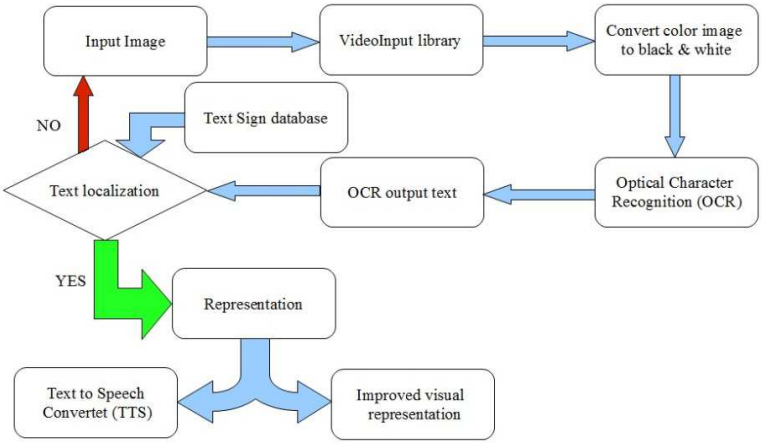


Figure 1. Process diagram

3.1 Capturing input image

First of all, the system must get the image from the required camera. If the system has several cameras, it is necessary to choose between them, keeping in mind some quality parameters, such as the camera resolution, the situation in the device (front or back), etc. As the system proposed is made up of a tablet with a webcam and an external high definition camera, the system will consider the camera with the highest resolution. To carry out this selection, the system takes advantage of the functions provided by the videoInput library (Web link: muonics.net/school/spring05/videoInput, last access Jan 2011) (written in C/C++ language).

3.2 Image treatment

Once the image from the camera is captured by the system, the image has to be converted into a B&W one. For this purpose, the OpenCV library (OpenCV, 2011) provides the suitable tools to perform a useful treatment. Firstly, the image must be converted to a gray scale. The CvtColor function makes the conversion from the RGB color space to a gray scale as follows: RGB to Gray: $Y \leftarrow 0.299 \bullet R + 0.587 \bullet G + 0.114 \bullet B$

In order to obtain a right conversion, the image type must be squared up (8-bit, 16-bit or floating-point image).

Figure 2. Color image (left) and gray scale image (right).

From this gray scale image, the system is able to generate the B&W image that it will be used in the following steps of the process. To obtain this B&W image, the Threshold function extracts a bi-level (binary) out of the previous gray scale image the way the threshold type parameter indicates. Considering that the text in sign can be represented in positive (letters are darker than background) or negative forms (letters are lighter than background), the proposed system uses both possibilities: binary and binary inverted type (input and output shown in Fig 4), that correspond to the following equations respectively:

$$dst(x,y)=\begin{cases} max.value & if\ src(x,y)>threshold \\ 0 & otherwise \end{cases}$$

$$dst(x,y)=\begin{cases} 0 & if\ src(x,y)>threshold \\ max.value & otherwise \end{cases}$$

In the formulas, dst stands for destiny and src for source. In our case, maximum value is the integer 255 (white in an 8-bit space color image, 0 is black). Finally, the threshold is the value that de-fines the border between the pixels which will turn black and those turning white. This threshold is controlled by a scrollbar (Figure 3) defined in the black and white window, being able to be adapted automatically depending on the ambient light.

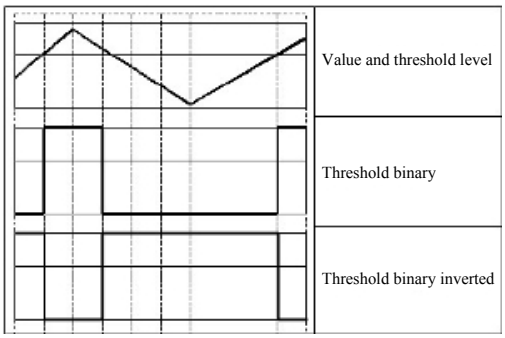


Figure 3. Input and output of the Threshold function

Figure 4. Scrollbar in the black and white window

3.3 Optical Character Recognition

The OCR used in the system is Tesseract (Web link: http://code.google.com/p/tesseract-ocr/ (last access Jan 2011)). Tesseract is open software that was primarily designed for pulling out the text from .pdf files.

This design is the reason for the previous image treatment, providing a similar image to those that the OCR is expecting. The OCR is expecting text images in a positive way: letters darker than back-ground.

Tesseract first reads the input file line by line searching for black pixels. This is the edge detection process, in which the OCR scans the input and splits black pixels into blobs that will be processed to extract outlines. This first part of the work made by the OCR is very interesting because it gives us the possibility to deal better with the light and shadows problem in the image capture. Sometimes it will be easier to obtain a white background and black letters from the sign and other times white letters in front of a black background will be clearer. In both cases, when the characters are wide enough, the outlines obtained after the edge detection are similar.

After the edge detection, the lines are adjusted for skew and the gradient or rotation of the text is calculated. Then, a high-level procedure segments the blobs into possible words (considering the spacial disposition of the blobs) and space them. Finally, a classification method converts the analyzed features into letters and makes a quality check, saving the result.

Although Tesseract can read a wide variety of image formats and convert them to text in over 40 languages, a low accuracy could be improved by training it. This is useful when the character type is uncommon or a clear image of the text is not being enough to obtain appropriate results.

Using the OCR could be a consuming process depending on the CPU used by the device. As our system is designed for portable devices, which usually have not very high-powered CPUs, it is not possible to repeat the whole process for all frames (25 per second). Otherwise, a possible user does not need this time resolution, analyzing one frame every second or every 0.5 seconds would be enough: everybody needs a minimum interval of time to change his or her position significantly. The time between text searches can be adjusted by the user by means of a scrollbar situated in the main window (Figure 5).

Figure 5. Scrollbar controlling time between text searches

When the OCR finishes, the text output is saved in a .txt file. This file is available for a limited time because it will be overwritten with the new results obtained by Tesseract in the following search.

3.4 Text Localization

When the OCR finishes its work, the system must look into the given file for any coincidence with one of the signs stored in the database. The OCR output (a string) must be compared to the text of the signs (text patterns) stored in the database. This comparison is made by using an algorithm inspired by the Levenshtein distance (Navarro, 2001; Woltzenlogel, 2007)

The Levenshtein distance is a way of measuring the distance between two strings. It is used by the correctors of the text processors, offering words with a maximum distance between them and the wrong written one. The main challenge faced by the proposal presented in this paper is the way of dealing with the noise that the OCR includes when it finds some patterns in the image similar to a character. Same examples can be seen in figures 6 and 7.

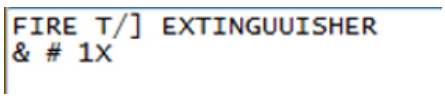


Figure 6. Output file when processing an ‘FIRE EXTINGUISHER’ sign

Figure 7. Output file when processing a 'EXIT' sign

As is shown, sometimes there are too many mistaken characters introduced by the OCR. That is why in this case the Levenshtein distance is not useful in its original form. To fix this problem, the proposed algorithm compares the text contained in the file handled by the OCR in a different way with every sign saved in the database. When comparing one, it looks for the first character of the sign in the file text. If there is a hit, it will look for the next character of the sign text starting now from the following position which was successful. If not, it will look for the next character from the last starting position. The algorithm saves the sign which has the highest number of hits only in case the number of hits divided by the number of characters contained in the text sign exceeds a defined and tunable percentage.

Figures 8, 9 and 10 illustrate this process, where the input is EMERGENCY EXIT.

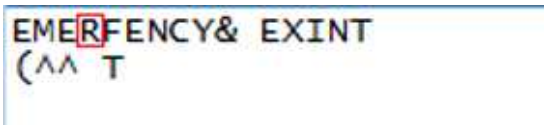


Figure 8. Hit on character 'R'

As shown in Figure 8, there is a hit on the character 'R', so the algorithm is going to look for the character 'G' (the following one in the sign text) starting from the position where character 'F' is situated in the file, as shown in figure 9:

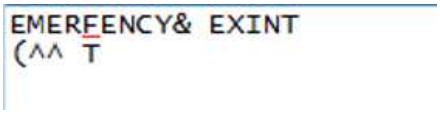


Figure 9. Starting position for the next search during comparison

The algorithm is not going to find any 'G' character among the remaining characters. From this position, the next hit will be with character 'E':

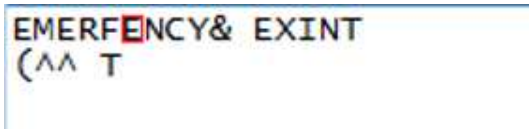


Figure 10. Hit on 'E'

When the algorithm ends, two results are reported:

- 13 hits (spaces are included)
- 0.62 (13 hits over 21 characters in the signs) ratio (62 %).

If the percentage (developing tuning) is enough, and the number of hits is the highest one, this signal will be represented.

Also, in this example, it is possible to see the importance of the number of hits. The sign EMERGENCY EXIT also contains the sign EXIT. If the success ratio were the only controlling variable, as EXIT is contained in EMERGENCY EXIT and also being shorter, it is more likely to have a higher ratio (percentage). In this case, the system would propose EXIT as the sign detected by the system, thus producing an error. Nevertheless, by using the number of hits as a controlling variable, the system may be able to distinguish examples similar to this.

3.5 Representation

Once a sign is detected, it must be presented for the user. Considering possible users with different levels of visual impairment, the system will provide two different representations. First of all and thinking of blind people, the main presentation must

be an audible one by converting the text into speech. For this purpose, there are several Text to Speech systems available. The choice of one of them is not so relevant because any of them will read any given text sign. The differences could be found during the inclusion in the system (some of them could be easier to include). In this work, the Loquendo (Web link: http://www.loquendo.com (last access Jan 2011)) system was integrated.

Figure 11. System output during representation

The second representation is a visual improved output. The text detected by the system will be printed out on the screen in a bigger and brighter format, helping the users with visual difficulties. After the representation, the complete process will finish and start again with a new frame (extracted from the video signal) after a certain time (1 or 0.5 seconds depending on the user configuration). Figure 11 shows a capture of the complete system when representing the EXTINTOR sign information (audio is assumed but not represented in the figure).

4 Conclusions

This paper has presented a low complexity strategy for detecting and recognizing text signs automatically. Instead of using lengthy image algorithms for detecting the text sign, and then, applying an Optical Character Recognition (OCR) algorithm in the areas previously identified, this paper proposes the application of the OCR in a whole lightly treated image and then carries out the text detection process on the OCR output.

This paper has also described the implementation of a real time system. The system is made up of five main modules: the first one captures the image (or sample) from a video camera. Secondly, the color image is converted into a gray scale and then into a black & white image. The B&W image is passed to the OCR that will provide a text string. The text localization process is carried out on this text string considering several text patterns (from the most relevant text signs) that have been stored in a database. Finally, the text pattern detected is converted into speech for blind people and into a bigger and brighter text for the visually impaired.

5 Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement n° 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), ITALI-HA (CAM-UPM), INAPRA (MICINN, DPI2010-21247-C02-02), SD-TEAM (MEC, TIN2008-06856-C05-03) and MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542) projects. Authors also thank Mark Hallett for the English revision of this paper and all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics.

6 References

X. Chen, J. Yang, J. Zhang, and A. Waibel, "Automatic detection and recognition of signs from natural scenes," IEEE Trans. Image Process., vol. 13, no. 1, pp. 87–99, Jan. 2004.

X. Huang and H. Ma, "Automatic detection and localization of natural scene text in video," in Proc. 2010 20th Int. Conf. Pattern Recognition (ICPR), 2010, pp. 3216–3219.

W. Jirattitichareon and T. H. Chalidabhongse, "Automatic detection and segmentation of text in low quality Thai sign images," in Proc. IEEE Asia Pacific Conf. Circuits and Systems, 2006, pp. 1000–1003.

K. Jung, J. Han, K. Kim, and S. Park, "Support vector machines for text location in news video images," in Proc. TENCON, 2000, vol. 2, pp. 176–180.

K. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1631–1638, Dec. 2003.

G. Navarro (2001). "A guided tour to approximate string matching". ACM Computing Surveys 33 (1): 31–88

OpenCv Library http://www.seas.upenn.edu/~bensapp/opencvdocs/ref/opencvref_cv.htm (last access Jan 2011)

P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian approach to multi-oriented text detection in video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 412–419, Feb. 2011.

B. Sin, S. Kim, and B. Cho, "Locating characters in scene images using frequency features," in Proc. IEEE Int. Conf. Pattern Recognition, 2002, vol. 3, pp. 489–492.

B. Woltzenlogel Paleo. An approximate gazetteer for GATE based on levenshtein distance. Student Section of the European Summer School in Logic, Language and Information (ESSLLI), 2007.

W. Wu, X. Chen, and J. Yang, "Detection of text on road signs from video," IEEE Trans. Intell. Transp. Syst., vol. 6, no. 4, pp. 378–390, Dec. 2005.

J. Yu and Y.Wang, "Apply SOM to video artificial text area detection," in Proc. Int. Conf. Internet Computing in Science and Engineering, 2009, pp. 137–141.

TexAFon: a multilingual text processing tool for text-to-speech applications

Juan María Garrido[1], Yesika Laplaza[2], Montserrat Marquina[2], Claudia Schoenfelder[2], Sílvia Rustullet[1]

[1]Computational Linguistics Group (GLiCom), Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona
juanmaria.garrido@upf.edu, silvia.rustullet@upf.edu
[2]Speech and Language Group, *Barcelona Media Centre d'Innovació*, Barcelona
yesika.laplaza@barcelonamedia.org,
montse.marquina@barcelonamedia.org,
claudia.schoenfelder@barcelonamedia.org

Abstract. This paper describes TexAFon, a set of multilingual text processing tools for automatic normalization, phonetic transcription, syllabication, prosodic segmentation and stress prediction from text. The paper describes the general architecture of the system, its main modules, the obtained output and its current applications. It outlines also the ongoing improvements for next version.

Keywords. Text-to-speech, text processing, Spanish, Catalan.

1 Introduction

In this paper we present TexAFon, a set of Catalan/Spanish text processing tools for automatic normalization, phonetic transcription, syllabication, prosodic segmentation and stress prediction from text. It has been conceived to perform all the linguistic processing usually applied to the input text in text-to-speech systems, but it can also be used for other purposes, such as automatic phonetic transcription of texts or automatic building of phonetic dictionaries. It has been jointly developed by researchers of the Computational Linguistics Group (GLiCom) of Pompeu Fabra University and the Speech and Language Group at Barcelona Media.

Fully developed in Python [1], linguistic knowledge in TexAFon has been implemented in the form of:

- Python procedures, containing the linguistic rules;
- Python lists, containing non-editable information;
- External dictionaries, stored in text files (then editable by external users).

TexAFon has a modular architecture (Figure 1), which facilitates the development of new applications using it, the addition of new languages, and the connection to other external modules and applications. This architecture clearly differentiates among:

- A general processing core, which includes the language-independent procedures.
- The language packages (two, for Spanish and Catalan), including modules and dictionaries specific of the language.
- The applications, which call the processing core depending on their needs.

Next sections present the structure and workflow of these components in more detail.

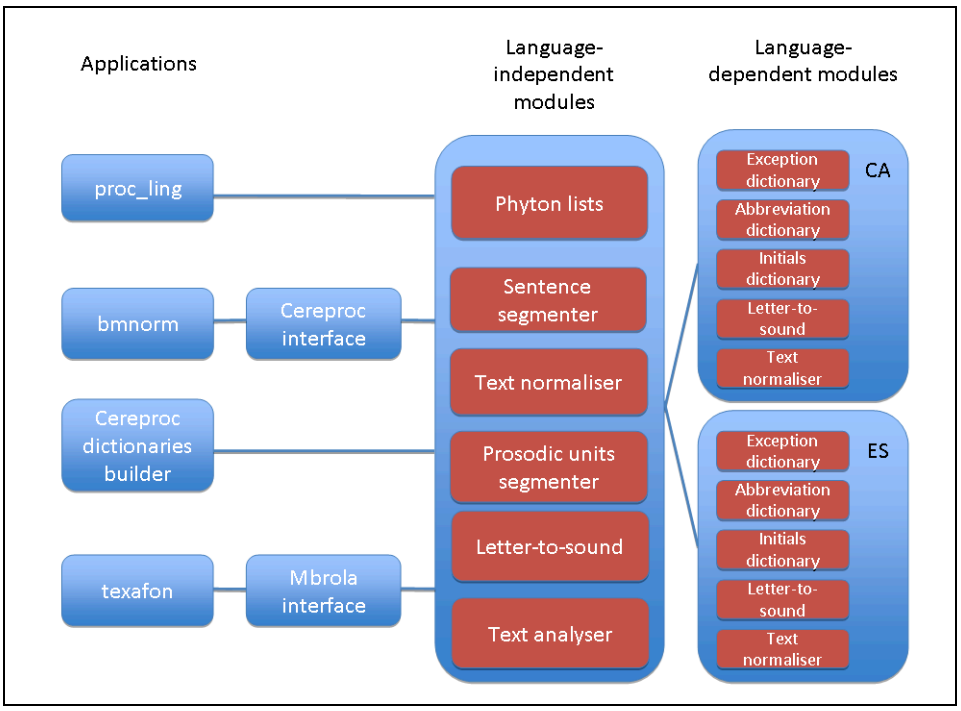


Fig. 1. General TexAFon architecture

2 The processing core

The processing core of TexAFon includes a set of language-independent Python modules, each one responsible of a different text processing task:

- ***Segmenta_frases***: in charge of the segmentation of the input text into sentences.
- ***Textnorm***: responsible of the expansion of numbers, symbols, abbreviations and any other 'irregular' expression appearing in the input text.
- ***Segmentacion_prosodica***: in charge of segmenting the input text into prosodic units.
- ***Lts***: in charge of phonetic transcription, syllabication and stress prediction.
- ***Analisis_texto***: in charge of extracting relevant linguistic/paralinguistic information (in the current version, only sentence mood) from the input text.

Depending on the selected language (Catalan or Spanish), these modules call the corresponding language-dependent modules, which make use of the corresponding language-dependent dictionaries, as shown in figure 1.

2.1 Sentence detection

Segmenta_frases is the module in charge of segmentation of the input text into sentences. This module is language-independent, and basically looks for punctuation marks indicating the end of a sentence (dots, question and exclamation marks). Ambiguous cases, such as dots appearing at the end of abbreviations or initials ('*J. L. García*'), are treated properly.

2.2 Text normalisation

Text normalisation is carried out in TexAFon by the *textnorm* module. This task usually involves the conversion of non-orthographic text elements (such as e-mail addresses, dates, abbreviations, etc.) to their corresponding fully-orthographic form. *Textnorm* performs normalisation in two steps:

- **Labelling**: every token (chain of characters between two blank spaces) of the text is assigned to one of the categories included in table 1.
- **Expansion**: every token is processed properly, according to its label, to obtain its corresponding readable full orthographic version.

Labelling is performed mostly by rule, by calling language-independent procedures which analyze the form of the token and assign it to a category. So, por example, the token '*11/06/2012*' is labelled as '*fecha*' because it is made up of a series of figures and slashes that fit the usual form of a date. In some cases, labelling procedures make use of language-dependent dictionaries (which can be modified by the user) or lists (non-editable) to check if a token belongs to a particular category, as in the case of abbreviations ('*Sr.*').

Expansion is carried out by language-dependent procedures, which return the expanded form corresponding to the selected language. Special attention has been paid to the expansion of tokens made up of combinations of letters, figures and symbols ('*B-737*'; '*1-1*', '*22@*'), which usually are not properly expanded in similar text processing modules. Words (or word-like forms) showing a non-regular syllabic structure receive also a special treatment: if they are not found in the exceptions dictionary, they are split into sets of 'character islands' that make them readable by a TTS. This is especially useful, for example, in the case of e-mail or internet addresses, whose pronunciation may be difficult for a TTS.

Table 1. Token categories considered in pre-processing

palabra	*porcentaje*	*numero_entero*	*numero_decimal*
abreviatura_ordinal	*fecha*	*hora*	*numero_romano*
numero_telefono	*cifras*	*direccion_electronica*	*sigla*
abreviatura	*cadena_con_simbolos*	*cadena_con_signos*	*cifras_letras_y_signos*
cifras_y_letras	*cifras_y_signos*	*letras_y_signos*	*signo*
simbolo	*temperatura*	*letra_aislada*	*expresion_monetaria*

2.3 Prosodic unit detection

Segmentacion_prosodica is the module in charge of segmenting the input text into prosodic units. These units are relevant, in text-to-speech systems, for the prediction of prosodic parameters, such as the location and length of pauses, phone durations, and F0 contours. Prosodic segmentation is actually done in two steps:

- **segmentation of text into Intonation Groups (IG)**, which is carried out before phonetic transcription;
- **Stress Group (SG) detection**, which is carried out after phonetic transcription.

Both units, IG and SG, are currently used by GenProso, the F0 prediction module developed at GLiCom, which assumes the F0 modelling framework described in [2], [3], [4], and which is used in conjunction with TexAFon by some applications.

Intonation Group detection. IG is the basic unit for the prediction of F0 contours in GenProso: each IG contains a single F0 contour. It is also the reference unit for pause insertion in the current implementation of TexAFon. IG detection consists basically of the detection of punctuation marks associated with the realisation of a pause in reading, such as dots, comma, parentheses, etc. ('orthographic' pauses). They are detected in the text, and then substituted by their corresponding pause label. 'Non-orthographic' pauses (those not related to punctuation marks, and only predictable if a morpho-syntactic parser is available) are not predicted at this moment.

Stress Group detection. SG is also an important unit for the building of F0 contours in GenProso. GA are defined as sets of syllables made up of a stressed syllable and all the unstressed syllables following it, until the next stressed syllable or the end of the IG. Unstressed syllables before the first stressed one in a IG are included in the first SG. SGs are detected in TexAFon once the syllable segmentation is available, that is, after phonetic transcription. The input for this process is the chain of phone symbols and syllable boundaries, which is segmented into IG by detecting the stressed syllables. The output of this module is a chain of SG for each of the IG of the input text.

2.4 Phonetic transcription, stress prediction and syllabication

Lts is the TexAFon module responsible of phonetic transcription and stress prediction, two typical tasks in text-to-speech. It is also in charge of syllable detection.

Catalan and Spanish orthographies are quite regular. For this reason, transcription and stress prediction tasks are performed in TexAFon by a set of language-dependent contextual rules: in the case of transcription, the output transcription symbol corresponding to the input character is predicted as a function of their previous and next neighbor characters; stress assignment rules predict the stressed vowel within a word as a function of the position of the stress mark (if any) in the orthographic form of the word. In both cases, exceptions (mainly foreign words and proper names) are handled by means of language-specific exception dictionaries.

Lts module works in two steps:

- **'Word-by-word' processing**: input text is processed word-by-word, in order to obtain word phonetic transcription, stressed syllable prediction and syllabication. Each word in the input text is first checked in the corresponding exception dictionary. If found, transcription and stress prediction are directly taken from the dictionary. Then, irrespective of the source of the transcription (rule or dictionary), syllabication is carried out.
- **IG processing**: the output of the previous step, containing phonetic transcription, stress marks and word/syllable boundaries, is processed at IG level to take into account possible phonetic phenomena occurring between word boundaries. Context-dependent phonotactic rules are used at this point.

Lts can generate the output transcription in two different formats: SAMPA [5] and the one use by Cerevoice®, the TTS engine by Cereproc [6]. Transcription includes word (#) and syllable (.) boundaries, in addition to the phone and stress symbols.

2.5 Text analysis

Analisis_texto is the module responsible of predicting from the text a series of relevant linguistic information. Currently only sentence mood is predicted: input sentences are labelled as declarative, interrogative or exclamative depending of the punctuation mark appearing at the end of the sentence. In the case of interrogative sentences, further processing is made in order to differentiate between total and partial questions.

3 Language packages

The processing core makes use of the necessary language-dependent resources to process the input depending on the chosen language, as shown in figure 1. TexAFon has currently two different language packages ('ca' and 'es'), both containing:

- Two language-dependent modules:

  - an *lts* module, containing the language-specific functions for letter-to-sound translation, stress prediction and syllabication;
  - A *textnorm* module, containing the language-dependent functions for text expansion.
- Three dictionaries:
  - an exceptions dictionary, containing the transcription of words which are not regular;
  - an abbreviations dictionary, containing the most frequent and general abbreviations in the given language, and their corresponding expansion;
  - an initials dictionary, containing initials whose pronunciation do not follow the general spelling rules.

4 Output

TexAFon provides then different levels of processed output for a given input text, both in Spanish and Catalan:

- segmentation into sentences;
- segmentation into IG;
- full normalisation and expansion;
- phonetic transcription (SAMPA or Cereproc);
- syllabication;
- segmentation into SG;
- sentence mood labelling.

These different outputs can be used separately by the final applications depending on their needs. Table 2 shows an example of these outputs: for the input text in (1), TexAFon provides the outputs in (2, 3, 4, 5).

Table 2. Outputs produced by TexAFon for a given input text

(1) Input	Esta es una demostración de TexAFon para IberSpeech 2012. TexAfon ha sido desarrollado conjuntamente por el Grupo de Lingüística Computacional (GLiCom) de la UPF y el Grupo de Voz y Lenguaje de Barcelona Media. Hoy es 11/06/2012, y son las 10:15.
(2) Segmentation into sentences	Esta es una demostración de TexAFon para IberSpeech 2012. TexAfon ha sido desarrollado conjuntamente por el Grupo de Lingüística Computacional (GLiCom) de la UPF y el Grupo de Voz y Lenguaje de Barcelona Media. Hoy es 11/06/2012, y son las 10:15.
(3) Segmentation into IG	Esta es una demostración de TexAFon para IberSpeech 2012 [signo_punto] TexAfon ha sido desarrollado conjuntamente por el Grupo de Lingüística Computacional [signo_abre_parentesis] GLiCom [signo_cierra_parentesis] de la UPF y el Grupo de Voz y Lenguaje de Barcelona Media [signo_punto] Hoy es 11/06/2012 [signo_coma] y son las 10:15 [signo_punto]
(4) Normalisation	esta es una demostración de texafon para iberspeech dos mil doce [signo_punto] texafon ha sido desarrollado conjuntamente por el grupo de lingüística computacional [signo_abre_parentesis] glicom [signo_cierra_parentesis] de la u pe efe y el grupo de voz y lenguaje de barcelona media [signo_punto] hoy es once de junio de dos mil doce [signo_coma] y son las diez quince [signo_punto]
(5) Phonetic transcription	e_&quot s . t a # e_&quot s # u_&quot . n a # d e . m o s . t r a . T j o_&quot N # d e # t e k . s a_&quot . f o N # p a . r a # i . b e r s . p e . e_&quot tS # d o s # m i_&quot l # d o_&quot . T e [signo_punto] t e k . s a_&quot . f o N # a_&quot # s i_&quot . d o # d e . s a . rr o . L a_&quot . d o # k o n . x u_&quot n . t a . m e_&quot n . t e # p o r # e l # g r u_&quot . p o # d e # l i n . g w i_&quot s . t i . k a # k o m . p u . t a . T j o . n a_&quot l [signo_abre_parentesis] g l i . k o_&quot m [signo_cierra_parentesis] d e # l a # u # p e_&quot # e_&quot . f e # i # e l # g r u_&quot . p o # d e # b o_&quot T # i # l e n . g w a_&quot . x e # d e # b a r . T e . l o_&quot . n a # m e_&quot . d j a [signo_punto] o_&quot j # e_&quot s # o_&quot n . T e # d e # x u_&quot . n j o # d e # d o s # m i_&quot l # d o_&quot . T e [signo_coma] i # s o_&quot N # l a s # d j e_&quot T # k i_&quot n . T e [signo_punto]

5 Applications

TexAFon processing core provides then a set of modules to perform different text processing tasks and to obtain different outputs depending on the needs of the final applications. Its modular architecture allows the development of such applications with a minimum effort.

Several applications which make use of TexAFon have been developed already, which are currently being used both at Barcelona Media and GLiCom for different purposes:

- ***Proc_ling***: a general-purpose text processing tool for Catalan and Spanish;
- ***Bmnorm***: a linguistic processing module for the Cerevoice TTS engine;
- ***Genera_Cereproc_dicc***: a tool for the generation of phonetic dictionaries in the format needed by the Cereproc Voice Creation Kit;
- ***Mbrola***: a linguistic processing module providing input for the Mbrola TTS engine.

Proc_ling is a text processing tool for Spanish and Catalan which gives separately (stored in several text files) different outputs for the provided input text:

- segmentation into sentences;
- segmentation into IG;
- full normalisation and expansion;
- phonetic transcription and syllabication;
- segmentation into SG.

Bmnorm can be pipelined to the Cerevoice® TTS engine to produce synthetic speech in both both Spanish and Catalan, using the bilingual voices developed at Barcelona Media. It accepts as input both plain text and xml files following the Cerevoice® tagging conventions, and provides as output xml 'spurts' containing the linguistic information necessary to generate synthetic speech, in the format required by the Cerevoice® TTS engine. A special *Cereproc_interface* was developed to handle properly both Cerevoice input and output formats.

Genera_Cereproc_dicc is a development tool used at Barcelona Media for the building of phonetic dictionaries compatible with Cereproc Voice Creation Kit; it can be easily adapted to generate phonetic dictionaries in any other format.

Finally, *Mbrola* generates '.pho' files which serve as input to the Mbrola synthesis engine [7]. To do this, an *Mbrola_interface* has been developed. It makes use of GenProso, the prosody prediction tool developed at GLicom, to predict duration and F0 values for each output phone. It is being used at GLiCom as a research tool to investigate prosodic modelling for TTS purposes.

6 Ongoing improvements

A new version of TexAFon is expected to be available by the end of this year, including new functionalities especially oriented to the generation of expressive and domain-specific speech:

- **Normalisation of non-standard text**: special normalisation procedures are being included to convert input text written in non-standard form, such those of SMS or chats, to their corresponding standard orthographic form.
- **Emotion detection**: a procedure to categorise input sentences with emotional labels is still under development; this will be especially useful to investigate the automatic generation of emotional speech from untagged text.
- **Speech act detection**: another procedure is being developed to detect speech act in the input text; this will be especially useful to investigate the prediction of prosody in dialogue situations.

It is not expected to make a version of TexAFon available for public downloading in the near future, although authors are open to allow its use for specific projects and collaborations (including, for example, its connection to other speech synthesis engines).

7 References

1. Python Language Website, http://www.python.org/.
2. Garrido, J. M.: Modelling Spanish Intonation for Text-to-Speech Applications. Ph.D. thesis, Universitat Autònoma de Barcelona (1996)
3. Garrido, J. M.: La estructura de las curvas melódicas del español: propuesta de modelización. Lingüística Española Actual, XXIII/2, 173-209 (2001)
4. Garrido, J. M.: A Tool for Automatic F0 Stylisation, Annotation and Modelling of Large Corpora. In: Speech Prosody 2010, Chicago, May 2010. (2010)
5. Wells, J. et al.: Specification of SAM phonetic alphabet (SAMPA). In: A J Fourcin, G Harland, W Barry, and V Hazan, editors, *Speech Input and Output Assessment: Multi-lingual Methods and Standards.* Ellis Horwood Ltd Chichester (1989)
6. Garrido, J. M., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M., Pidcock, Ch.: The CEREVOICE speech synthesiser. In: Actas de las V Jornadas de Tecnología del Habla (Bilbao, 12-14 noviembre 2008), pp. 126-129 (2008)
7. Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vrecken, O.: The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: Proc. ICSLP 96, Philadelphia, 3, pp. 1393-1396 (1996)

The First European Portuguese Elderly Speech Corpus

Annika Hämäläinen[1,2], Fernando Miguel Pinto[1], Miguel Sales Dias[1,2], Ana Júdice[1], João Freitas[1], Carlos Galinho Pires[1], Vitor Duarte Teixeira[1,2], António Calado[1], Daniela Braga[1,2]

[1]Microsoft Language Development Center, Lisbon, Portugal
[2]ADETTI – ISCTE, IUL, Lisbon, Portugal

{ t-anhama, a-fpinto, Miguel.Dias, a-anjudi, i-joaof, a-capire, t-vitote, v-antonc, dbraga } microsoft.com

Abstract. Currently available speech recognisers do not usually work well with elderly speech because the underlying acoustic models have typically been trained using speech collected from younger adults only. To develop speech-driven systems capable of successfully recognising elderly speech, a sufficient amount of elderly speech data is needed for training the acoustic models. This paper describes an elderly read speech data collection effort carried out in Portugal, from the autumn of 2010 until the spring of 2012, applying a low-cost proctored data collection methodology using our in-house data collection platform, *Your Speech*. The resulting corpus of Portuguese elderly speech contains almost 90 hours of pure speech (speech with silences excluded). The transcriptions for about half of the data have already been manually verified and annotated for noises and disfluencies.

Keywords: Speech corpus, elderly speech, read speech, Portuguese

1 Introduction

Statistics from the 2001 census in Portugal [1] show that the Portuguese population is aging. This is mainly because of increasing life expectancy and a decreasing birth rate. In December 2010, the elderly (people aged 65 and over) already covered 18.2% of the Portuguese population – an increase of 1.1 percentage points as compared with 2005 [2]. The projections presented in [3] show that, by 2060, the elderly population is expected to increase by a figure between 88.2% (optimistic scenario) and 103.74% (pessimistic scenario). Similar demographic evolution is expected in many other European countries [4]. According to the World Health Organization [5] and considering the above statistics, it is of paramount importance that policies and programs to keep the older population active and productive are implemented. In fact, to provide equal opportunities for all, ensuring active aging is one of the biggest challenges that our society will have to face in the coming decades.

Previous research suggests that elderly people have more difficulties using information and communication technology (ICT) than younger adults do [6, 7]. The main reasons for this are the complexity of the existing user interfaces and the limited set of available interaction modalities caused by the technology mainly being designed with younger users in mind. Hence, from the point of view of ICT, ensur-

ing active aging translates into adapting the technology to better suit the needs of the elderly, for instance, by increasing the choice of available interaction modalities.

Our usability evaluation studies suggest that speech is the easiest and most natural modality of human-computer interaction (HCI) [8]. Speech is also the preferred modality when interacting with mobile devices when permanent user impairments like arthritis, or temporary limitations that might occur in "on-the-go" situations such as driving, make it difficult to use other modalities like touch. However, speech recognisers do not yet work well with elderly speech. This is because age influences recognition performance due to several parameters of the speech signal (e.g. fundamental frequency, jitter, shimmer and harmonic noise ratio) changing with age [9, 10] and because the acoustic models needed to recognise speech are usually trained using speech collected from the younger generations only. [11] showed that, as compared with middle-aged adults, the increase in word error rate (WER) can be as high as 57% relative when recognising elderly speech (in this case, digits read out by Danish elderly subjects aged over 70) using acoustic models trained with speech collected from young to middle-aged adults. On the other hand, in [12], recognising American English elderly speech (both spontaneous and read) using elderly-specific acoustic models led to a significant decrease (about 12.5% absolute) in WER as compared with acoustic models trained using speech collected from younger adults.

To build speech-driven systems capable of successfully recognising elderly speech, a sufficient amount of elderly speech data is needed for training the acoustic models. There is growing international interest in improving elderly speech recognition, with several past and ongoing initiatives to collect large corpora of elderly speech. JASMIN-CGN [13], resulting in a corpus of 25 hours of Dutch and Flemish elderly speech, and S-JNAS [14], resulting in a corpus of 133 hours of Japanese elderly speech, are examples of past initiatives. The goal of this paper is to describe the results of a two-year nationwide data collection campaign carried out by Microsoft Language Development Center (MLDC) to collect clean wideband speech read out by elderly Portuguese speakers – that is, 16-bit, 16/22-kHz elderly read speech collected in a quiet environment using a laptop and a USB microphone headset. In addition to our effort, Portuguese elderly speech data is only being collected by INESC-ID [15]; a collaboration between the two efforts is already under way.

2 Corpus Design

2.1 Speaker Selection

It is possible to observe differences between teenager speech, younger adult speech and elderly speech at the acoustic-phonetic level [16, 17]. However, there do not seem to be any clear-cut age-related acoustic-phonetic differences in speech [18]. This is partly due to the aging of speech also being influenced by factors like the abuse or overuse of the vocal folds, smoking, alcohol consumption, and psychological stress and tension [19, 20]. Furthermore, in Fiehler's opinion [21], what is usually considered typical of elderly speech is more related to situational circumstances, such as lexical and grammatical aspects, that are commonly used to identify different registers. While it might be impossible to determine the exact age from which an individual's speech could be considered elderly, studies usually regard 60-70 years

of age as the minimum age range for elderly speech [11]. Therefore, for our corpus, we decided to collect speech from subjects aged 60 and over.

A pilot study carried out by our team suggests that elderly people with a low level of education (e.g. primary school) might have poor reading and technical comprehension skills, and would therefore be difficult to collect read speech from [16]. Hence, we decided to concentrate our data collection efforts on elderly subjects recruited from universities for senior citizens where they would be guaranteed to have basic reading and technical comprehension skills. We collected speech from elderly subjects enrolled to 29 such universities in the Lisbon, Oeiras, Cascais and Porto areas of Portugal.

Apart from the age, literacy and basic technical comprehension requirements, we had no criteria for selecting the speakers. We did not, for example, aim at a specific ratio of female and male speakers or speakers from different regions.

2.2 Prompt Selection

The goal of the prompt selection was to create a set of prompts that would result in a corpus of read speech suitable for training acoustic models for a wide variety of speech-driven applications ranging from command and control through to dictation. The prompt selection and structure was inspired by existing Portuguese corpora such as SpeechDat II [22]. The prompts included phonetically rich sentences, as well as common types of prompts used in speech-driven applications (e.g. isolated digits, numbers, and date and time expressions). In addition to that, prompts specific to the speech-driven applications developed in the QREN Living Usability Lab (LUL) project [23] (e.g. the Living Home Center [24]) were included. Table 1 illustrates the number of unique speaker prompts selected for each prompt category.

Table 1. The number of different prompts for each prompt category.

Prompt Category	#Unique Speaker Prompts
Phonetically Rich Sentences	3497
Natural Numbers	750
Isolated Digits	10
Phone Numbers	150
PIN Codes	150
Dates	750
Times	1125
Time Ranges	370
Person Names	750
City Names	500
Application Words	229
Application Phrases	416
Web Words	29
PIM	210

The phonetically rich sentences were extracted from classic Portuguese literature and the CETEMPúblico corpus of Portuguese newspaper language [25]. To take the special needs of the targeted age group into account, long, complex sentences were

avoided, and sentences related to sensitive topics, such as death, were filtered out [17]. The natural numbers were decimal numbers containing 4-6 digits (e.g. *3,455).* In practice, speakers sometimes uttered them in unexpected ways. For instance, *579,327* might have been read out as the two cardinal numbers *579* and *327* separated with the word *virgula* ('comma'). Person names consisted of combinations of Portuguese first and last names (e.g. *Alice Neto*), and city names contained a total of about 500 Portuguese city names (e.g. *Aveiro*) and international city names in Portuguese (e.g. *Varsóvia* ('Warsaw')). Application phrases contained full commands considered relevant for the LUL applications (e.g. speech-enabled email (e.g. *abrir mensagem* ('open message')), calendar (*activar alarme* ('activate alarm'))), application words were shorter versions of them (e.g. *responder* ('respond'), *favoritos* ('favourites')), and web words consisted of important web sites (e.g. *Hotmail*), applications (e.g. *Internet Explorer*) and key strikes (e.g. *enter*). PIM refers to commands that might be used in Personal Information Management applications with commands for meeting request management, calendar management etc. In the case of number expressions (e.g. natural numbers, dates), the speakers were presented with their unnormalised forms in an attempt to capture all the possible variation in uttering these types of expressions. For instance, *1 de Janeiro* ('1 January') could be uttered as *Um de Janeiro* ('One of January') or as *Primeiro de Janeiro* ('The first of January'). Therefore, the number of the different types of utterances elicited by the prompts was sometimes higher than the number of the different prompts presented to the speakers (cf. Table 1).

Each subject was asked to utter a set of 160 prompts selected across the 14 prompt categories so that all categories and prompts would be represented in a balanced way in the corpus. Uttering 160 prompts resulted in recording sessions lasting about 20 minutes – a session duration that was short enough for maintaining the attention of the speakers and resulted in about 6-7 minutes of pure speech.

3 Data Collection

The goal of the data collection effort was to collect about 100 hours of pure speech read out by elderly Portuguese speakers. This was to be done at as low a cost as possible (less than € 15,000) – a goal that was achieved by recruiting and training students and work seekers to collect the data (see Section 3.2) using a web-based data collection platform (see Section 3.1). In practice, the workflow was almost completely automated and little involvement was needed from the Microsoft data collection team to operate the campaign.

3.1 Data Collection Platform

For collecting our data, we used the web-based *Doar a Voz* data collection platform, which is based on the *Your Speech* data collection platform developed at Microsoft [26]. The essential difference between the two platforms is that *Doar a Voz* is localised for European Portuguese and specifically adapted for elderly users, for example, by using larger font sizes and by recalibrating the speech detection algorithm such that it allows longer pauses, which are characteristic of elderly speech [17]. The platform basically consists of an HTTP server and a client application that rec-

ords the audio using a Silverlight component in Internet Explorer. The technical requirements for using the platform include the Windows operating system, Internet Explorer 9, Silverlight 4 Runtime, a broadband Internet connection and a noise-cancelling Life Chat LX 3000 USB headset. The web interface consists of a sequence of web pages used to inform the speakers of the data collection and their rights as a speaker, to collect information about them (age, gender and the region where they had grown up), and to present them with the prompts to read out. Each utterance is recorded separately in wave file format, without compression. The recording starts when the *Gravar* ('Record') button is clicked and ends when the *Parar* ('Stop') button is clicked or, alternatively, in an upgraded version of the system (see below), when no more speech is detected. The recorded utterance is then uploaded to the web backend and checked for the presence of speech and, in an upgraded version of the system, also for clipping; if everything is OK, the user may proceed to the next prompt by clicking the *Próxima Frase* ('Next prompt') button or, if (s)he so wishes, rerecord the utterance by clicking the *Regravar* ('Rerecord') button. If the quality control is not passed, the user is requested to rerecord the utterance. As opposed to data collection efforts using expensive DAT recorders, offline Wizard of Oz -based dialog managers and suchlike (e.g. [13, 14]), our data collection platform allows us to have an inexpensive, portable data collection setup that only requires a laptop with the above mentioned technical requirements but makes it possible for us to record speech from any speaker anywhere in the world.

The first version of the platform was used to collect a 54% of the speech data using a sampling frequency of 16 kHz and a 16-bit resolution. At the beginning of 2012, the platform was enhanced by adding an end of utterance detection algorithm as an alternative way of ending the recording of an utterance, and by improving the data quality control by refining the speech detection algorithm and adding an audio clipping detection algorithm (cf. Section 4.3). The upgraded version of the platform was used to collect the remaining 46% of the speech data using a sampling frequency of 22 kHz and a 16-bit resolution.

3.2 Data Collection Procedure

The data collection procedure involved three key steps and three key groups of participants: (1) negotiating the details of the data collection with the participating universities for senior citizens, (2) recruiting and training the proctors that were responsible for supervising the training sessions, and (3) arranging the recording sessions to record the speech of the elderly subjects.

Four Microsoft employees were responsible for establishing the contact with the universities, presenting the project to those in charge, highlighting the benefits of speech-driven applications that could be built using elderly speech data, explaining the incentive scheme, and signing collaboration contracts. The universities taking part in the project helped us spread the word about the data collection during classes and schedule recording sessions with their students. As an incentive, we paid the universities a small amount of money for each group of 20 elderly students taking part in the recordings. The speakers did not receive a payment; their incentive was to contribute to the development of speech-driven applications for the elderly.

As mentioned earlier, we recruited and trained students and work seekers as proctors who were responsible for supervising the recording sessions. During the ~2

years of the data collection effort, we worked with a total of 18 proctors. We paid them a small amount of money for each completed recording session and, later on during the campaign, introduced a performance reward model that improved the data collection rates considerably: we paid them an extra amount of money for each completed session if they recorded a minimum of ten sessions per week. During the recording sessions, which took place in a silent room, the proctors were responsible for: (1) setting the microphone of the headset at a distance of 3 cm from the speaker's mouth, (2) entering information about the speaker (age, gender and region) on the appropriate web page, (3) requesting the speaker to speak at a normal speech rate, and (4) depending on the speaker in question, controlling the *Doar a Voz* platform, or teaching the speaker how to use the platform by guiding her or him through the first couple of prompts.

Table 2. The number of speakers per age group and area of Portugal. The "Unknown" group contains speakers aged 60 and over whose exact age group was not recorded.

Age/Area	Centre	South	North	Madeira	Azores	Abroad	Total
60-65	240	21	128	1	0	10	400
66-70	106	14	47	2	0	12	181
71-75	86	15	50	2	1	6	160
76-80	55	6	25	0	0	5	91
81-85	32	3	18	0	0	0	53
86-90	18	1	12	0	0	1	32
91-95	0	0	0	0	0	0	0
96-100	0	0	1	0	0	0	1
Unknown	45	1	18	0	0	4	68
Total	582	61	299	5	1	38	986

Table 3. The amount of pure speech collected and annotated for each prompt category. At the time of writing this paper, about 24,000 utterances still remain to be annotated.

Prompt Category	hh:mm:ss	hh:mm:ss/Annotated
Phonetically Rich Sentences	50:06:49	24:00:03
Natural Numbers	02:52:45	01:24:37
Isolated Digits	00:16:40	00:08:19
Phone Numbers	05:45:60	02:49:39
PIN Codes	02:51:42	01:23:44
Dates	07:34:28	03:41:51
Times	04:10:35	02:03:02
Time Ranges	02:04:44	00:59:44
Person Names	02:48:11	01:20:58
City Names	03:15:31	01:34:20
Application Words	02:03:24	01:00:11
Application Phrases	03:42:27	01:46:38
Web Words	01:01:13	00:30:28
PIM	00:56:30	00:27:17
Total	89:30:59	43:10:50

Table 4. The tags available for annotating the elderly speech corpus.

Tag	Meaning
<FILL/>	Filled pauses (e.g. “umm”, “er”, “ah”)
<NON/>	Non-human noises (e.g. mouse or keyboard noises, radio, TV, music)
<SPN/>	Human noises (e.g. coughs, audible breath, prompt echoes)
<UNKNOWN/>	Unintelligible and truncated words, unclear speech from non-primary speakers

4 Corpus

4.1 Overview of the Speech Data

During the data collection effort, we recorded read speech from 986 elderly subjects aged 60 and over – 714 (72.4%) of them female and 272 (27.6%) of them male. Table 2 presents information about the age distribution and the regional origin of the speakers. As one might expect, the number of subjects per age group decreases with age. Demographically speaking, the collected speech data might not be fully representative of the elderly population in Portugal (e.g. Madeira and Azores are underrepresented and women are overrepresented). However, we managed to collect speech from a much larger slice of the elderly population (986 speakers out of a total of 2 million elderly people) than any other elderly speech data collection efforts that we are aware of (e.g. S-JNAS collected speech from301 speakers out of a total of 40.5 million elderly people [14]).

We recorded about 190 hours of data, of which almost 90 hours is pure speech. In terms of word tokens and types, the data contains about 810,000 word tokens and about 10,400 word types. Table 3 presents the amount of pure speech for each prompt category, together with the amount of pure speech whose transcriptions have been verified and annotated (cf. Section 4.2).

4.2 Transcription and Annotation

The transcription and annotation of the corpus is still underway. The orthographic transcriptions of about 50% of the data have already been verified manually. Two Microsoft employees trained for the task are listening to the recordings and, when necessary, changing the prompts presented to the speakers such that they correspond to what the speakers have actually said. All utterances containing number expressions are verified as the corresponding speaker prompts contained the unnormalised forms of the number expressions (cf. Section 2.2). The rest of the verified data originates from the beginning of the corpus, with all utterances included regardless of their category (cf. Table 1). All orthographic transcriptions are written in capital letters. All punctuation (e.g. commas) appearing in the speaker prompts is automatically removed; the only exception is spelled letters, which are disambiguated from one-letter words using a trailing period (e.g. “*A.*”).

While verifying the orthographic transcriptions, the annotators are also annotating filled pauses, damaged words (e.g. false starts), unintelligible words, human noises (e.g. coughs) and non-human noises (e.g. music, mouse clicks) using the tags presented in Table 4. The annotation work is carried out using an in-house transcrip-

tion tool. At the time of writing this paper, we have about 24,000 utterances containing, for instance, unnormalised number expressions left to be verified and annotated. The remaining ~35% of the data will not be verified or annotated.

4.3 Data Quality Control and Consistency

As mentioned in Section 3.1, the data collection platform has an instant data quality control step. The first version of the platform checked that the recordings do indeed contain speech and, in addition to that, the upgraded version checked that the recordings do not contain any clipping. Recordings not containing speech or containing even a single sample of clipping were rejected. For consistency, we intend to run the clipping detection algorithm also on the data collected using the first version of data collection platform before releasing the final version of the corpus. In addition, we use forced alignment to identify potentially problematic utterances; utterances that cannot be aligned are likely to have problems in their orthographic transcriptions and/or their audio quality. If possible, the problems are fixed; if not, the utterances are removed from the corpus.

To identify errors in the orthographic transcriptions of words, as well as "new" words, we break the orthographic transcriptions into individual words, and check those words against a large Portuguese lexicon. Using this simple method, we are able to identify misplaced tags (e.g. *vi*<SPN/>*rgula* ('co<SPN/>mma')), words that need to be added into the lexicon for the purpose of automatic speech recognition (e.g. *Facebook*), and words with typos. Some of the typos can be corrected automatically (e.g. *irgula* is bound to refer to *virgula*), while ambiguous typos need to be checked against the recordings (e.g. *ete* could refer to several different words).

A known shortcoming of the corpus is the missing age group information for 68 of the speakers (cf. Section 4.1). This problem could have been prevented by automatically validating the age groups entered by the proctors but was, unfortunately, spotted only after the data collection campaign had already finished.

Finally, similarly to [27], we intend to validate a subset of the data by checking the correctness of the file names and formats, the completeness of the data, the consistency between the speech data and the metadata, and the quality of the signals and their transcriptions and annotations. The validation will be performed by native speakers that have not been involved in the transcription and annotation work.

5 Conclusions

Between October 2010 and May 2012, we collected almost 90 hours (about 810,000 word tokens) of pure read speech from 986 elderly Portuguese subjects. Thanks to a web-based data collection platform and the use of students and work seekers as proctors to supervise the recording sessions, we were able to complete the data collection at a very low cost (less than €15,000, without the cost of verifying and annotating the orthographic transcriptions). In fact, we consider our effort a successful example of how to collect a large corpus of high-quality read speech at a low cost from a subset of the population that is relatively difficult to engage.

Researchers at INESC-ID in Lisbon, with whom we are actively collaborating, have already used a subset of the corpus to train and test elderly-specific acoustic models [28]. At the moment, we are also training and testing elderly-specific acoustic models using several different kinds of training setups; the best models will be used in the Living Home Center prototype [24] and other speech-driven applications aimed at the Portuguese elderly. Having learnt from our experiences during the data collection effort, we are now in charge of similar elderly speech data collection efforts in France, Hungary and Poland. These efforts are carried out in the scope of the European Paelife project, which is coordinated by MLDC and aims at fighting the isolation of the elderly and at allowing them to remain more independent, productive, and socially engaged. The end goal of the project is to develop a Personal (Virtual) Life Assistant, a virtual presence at home that supports learning, entertainment and social communication [29]. In addition to the ongoing data collection efforts, we are planning a similar effort in the mobile domain in Portugal.

The corpus is available at request for R&D activities. Please contact Miguel Sales Dias (Miguel.Dias@microsoft.com) for further information.

Acknowledgements. This work, which has been carried out in the scope of the QREN 7900 Living Usability Lab project, is co-funded by Microsoft, the Portuguese Government, and the European Structural Funds for Portugal (FEDER) through COMPETE (Operational Program for Competitiveness Factors), as part of the National Strategic Reference Framework (QREN), the national program of incentives for Portuguese businesses and industry. The authors are indebted to the speakers, universities and proctors involved in the data collection effort.

References

1. Instituto Nacional de Estatística: CENSOS 2001: Análise da população com deficiência, resultados provisórios. Lisbon (2002)
2. Instituto Nacional de Estatística: Estatísticas demográficas 2010. Lisbon (2012)
3. Instituto Nacional de Estatística: Projecções de população residente em Portugal 2008-2060. Lisbon (2009)
4. European Union Center of North Carolina: The EU's demographic crisis. EU briefings (2008)
5. World Health Organisation: Active aging: A policy network. Second United Nations World Assembly on Ageing. Madrid (2002)
6. Teixeira, V., Pires, C., Pinto, F., Freitas, J., Sales Dias, M., Mendes Rodrigues, E.: Towards elderly social integration using a multimodal human-computer interface. In: Proc. International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications, AAL 2012 @ BIOSTEC 2012. Vilamoura (2012)
7. Sfyrakis, S.A., Stephanidis, C., Akoumianakis D., Sfyrakis M., Paramythis A., Crete, H.: Universal accessibility in HCI: Process-oriented design guidelines and tool requirements. In: Proc. ERCIM Workshop on "User Interfaces for All", pp. 19-21. Stockholm (1998)
8. Teixeira, A., Braga, D., Coelho, L., Fonseca, J., Alvarelhão, J., Martín, I., Queirós, A., Rocha, N., Calado, A., Dias, M.: Speech as the basic interface for assistive technology. In: Proc. International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion. Porto Salvo (2009)
9. Neto, N., Patrick, C., Adami, A.G., Klautau, A. G.: Spoltech and ogi-22 baseline systems for speech recognition in Brazilian Portuguese. In: Teixeira, A., Lima, V., Oliveira, L.,

Quaresma, P. (eds.): Proc. International Conference on Computational Processing of Portuguese 5190, 256-259. Springer, Heidelberg (2008)
10. Xue, S.A., Hao, G.J.: Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study. In: Journal of Speech, Language, and Hearing Research 46, 689-701 (2003)
11. Wilpon, J. G., Jacobsen, C. N.: A study of speech recognition for children and the elderly. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, p. 349. Atlanta (1996)
12. Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R.: Recognition of elderly speech and voice-driven document retrieval. In: Proc. Acoustics, Speech, and Signal Processing (1999)
13. Cucchiarini C., Van hamme, H., van Herwijnen, O., Smits, F.: JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In: Proc. International Conference on Language Resources and Evaluation, pp. 135-138. Genoa (2006)
14. Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K.: Elderly acoustic models for large vocabulary continuous speech recognition. In: Transactions of the Institute of Electronics, Information and Communication Engineers, D-II J85D-II, pp. 390-397 (2002)
15. R&D Projects – AVOZ – Models for automatic speech recognition for elderly, http://bit.ly/OvbraV
16. Júdice, A., Freitas, J., Braga, D., Calado A., Dias, M., Teixeira, A., Oliveira, C.: Elderly speech collection for speech recognition based on crowd sourcing. In: Proc. International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion. Oxford (2010)
17. Helfrich, H.: Age markers in speech In: Scherer, K. & Giles, H.: Social markers in speech. Cambridge University Press, Cambridge (1979)
18. Ryan, E., Cole, R.: Evaluative perceptions of interpersonal communication with elders. In: Giles, H., Coupland, N. & Wiemann, J. M. (eds.): Communication, health and the elderly, pp. 172-191. Manchester University Press, London (1990)
19. Linville, S.E.: Vocal aging. Singular, San Diego (2001)
20. Jessen, M.: Speaker Classification in forensic phonetics and acoustics. In: Müller, C. (ed.) Speaker Classification I. LNCS (LNAI), vol. 4343. Springer, Heidelberg (2007)
21. Fiehler, R.: Kommunikation im Alter und ihre sprachwissenschaftliche analyse. Gibt es einen kommunikationsstil des alters? In: Sandig, B. & Selting, M. (Hrsg.): Sprech- und Gesprächsstile, pp. 345-370. Berlin/New York: de Gruyter (1997)
22. SpeechDat, http://www.speechdat.org/SpeechDat.html
23. Living Usability Lab, http://www.livinglab.pt/
24. Pires, C., Pinto, F., Teixeira, V., Freitas, J., Dias, M.: Living Home Center – A personal assistant with multimodal interaction for elderly and mobility impaired e-inclusion. In: Proc. International Conference on Computational Processing of Portuguese. Coimbra (2012)
25. CETEMPúblico, http://www.linguateca.pt/cetempublico/
26. Freitas, J., Calado, A., Braga, D., Silva, P., Dias, M.: Crowd-sourcing platform for large-scale speech data collection. In: Proc. FALA 2010. Vigo (2010)
27. Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., Baayen, H.: Experiences from the Spoken Dutch Corpus project. In: González Rodrigué z, M. and Paz Suarez Araujo, C. (eds.): Proc. International Conference on Language Resources and Evaluation 1:340-347. Las Palmas de Gran Canaria (2002)
28. Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M., Braga, D.: Impact of age in ASR for the elderly: preliminary experiments in European Portuguese. In Proc. IberSPEECH 2012. Madrid (2012)
29. Paelife, http://paelife.wordpress.com/project/

Let's Speak Quechua: The Implementation of a Text-to-Speech System for the Incas' Language

John Edgar Vargas Muñoz, Juan Antonio Cruz Tello and Richard Alexander Castro Mamani

Universidad Nacional de San Antonio Abad del Cusco, Av. de la Cultura Nro. 733
Cusco, Perú
jonv147@hotmail.com, j.soccoo@gmail.com, richardtk_1@hotmail.com
WWW home page: http://www.unsaac.edu.pe/

Abstract. The extinction of indigenous languages is one of the most crucial concerns that our societies are facing nowadays. In Peru, there have been many efforts to protect them from disappearance but almost all of them are done without the support of ICT. In this paper, we describe the development process of a text-to-speech (TTS) system for Quechua, the most widely spoken indigenous language of the Americas [1], [2]. For this procedure, we used the Cluster Unit Selection Synthesizer of the Festival Speech Synthesis System. The new male synthetic voice for Quechua will find direct application as part of an assistive technology for visually impaired and illiterate end users. To our knowledge, this project is between the few ones making an attempt to develop a TTS system for an indigenous language in America and the first for Quechua.

Keywords: text-to-speech, system, indigenous, language, Quechua

1 Introduction

The TTS System for Quechua will be an advantage for Quechua speakers, specially for those who do not know how to read, or do not read well. According to table 1 and 2 which describe the illiteracy rate of Quechua speakers, we can observe that more than the 20% of the Quechua speakers do not know how to read or write. And it is just for these people that this system could be useful. This census was done by the INEI (Instituto Nacional de Estadística e Informática, Perú) in 2007.

The Festival Speech Synthesis System is written in C++, has a multi-lingual support, offers a free general framework for building speech synthesis systems for new voices, has a Scheme (SIOD) based command interpreter for control, and has become available for many operating systems like Linux and its many distributions, Mac, and Windows. We chose it for these reasons.

The organization of the paper is as follows. Section 2 discusses Quechua and its phonological aspects. Section 3 discusses the implementation process. Section

Table 1. Numbers of speakers and illiteracy [5], Peru

	1940	1972	1981	2007
Total Pop. Age+	5.227 Mill.	11.331 Mill.	14.573 Mill.	24.687 Mill.
Quechua in %	2,903 Mill.	3,602 Mill.	3,755 Mill.	3,261 Mill.
	55.5%	31.79%	25.77%	13.2%

Table 2. Numbers of speakers and illiteracy [5], Peru

	Total Pop. age 15+	Spanish	Quechua
	19,054,624	15,814,303	2,676,420
literate	17,695,066	15,140,289	2,092,977
illiterate	1,359,558	674,014	583,443
illiteracy rate	7.1%	4.3%	21.8%

4 discusses the results. Then in section 5 we discuss the conclusions.

2 Quechua

Quechua [4] is a group of closely related languages, spoken by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-West of Argentina. Ethnologue also lists some Quechua speakers for Chile. Quechua is one of the official languages of Peru and Bolivia. Peru especially, has increased efforts to provide its citizens with official information not only in Spanish, but also in Quechua and (to less extent) in some other indigenous languages like Aymara and Asháninka.
In fig. 1, we can observe that the Quechuan Languages are divided into two main branches, Quechua I and II in terms of the Peruvian linguist Torero (1964). And these two are subdivided into dialects named used the letters A-C that represent the linguistic distance.

Quechua [4] is an essentially agglutinative language this characteristic differentiates Quechua from the dominant languages in the Americas (mainly English, Spanish and Portuguese), and throughout the years it has struggled with strong social pressure from these dominant languages.

3 Implementation Process with Festival

The first step to build the TTS system is to define the phoneset, for this project we used the Quechua II-C spoken in the city of Cusco with the writing standard proposed by the "Academia Mayor de la Lengua Quechua". The phoneset inventory we used for TTS system was developed according to [6] and [7]. In table 3 and 4 we show these symbols.

Table 3. Vowels and their phonetic representations as they were used in the TTS System for Quechua

GRAPHEME	IPA	SAMPA	VOWEL LENGTH	VOWEL HEIGHT	VOWEL FRONTNESS	LIP ROUNDING
A	a	a	Short	Low	Front	Unrounded
E	ɛ	E	Short	Mid	Front	Unrounded
I	I	I	short	High	Front	Unrounded
O	ɔ	O	Short	Mid	Back	Rounded
U	u	U	Long	High	Back	Rounded

Table 4. Consonants and their phonetic representations as they were used in the TTS System for Quechua

GRAPHEME	IPA	SAMPA	TTS Phonology Symbols	Consonant Type	Place of Articulation	Consonant Voicing
CH	tʃ	tS	ch	Affricative	Palatal	Voiceless
CHH	tʃʰ	tS_h	ch	Affricative	Palatal	Voiceless
CH'	tʃ'	tS_>	chx	Affricative	Palatal	Voiceless
H	h	h	h	Fricative	Post-velar	Voiceless
K	k	k	k	Simple Stop	Velar	Voiceless
KH	kʰ	k_h	kh	Aspirated Stop	Velar	Voiceless
K'	k'	k_>	kx	Glottalized Stop	Velar	Voiceless
L	l	l	l	Lateral Stop	Alveolar	Voiced
LL	ʎ	L	L	Lateral Stop	Palatal	Voiced
M	m	m	m	Nasal Stop	Bilabial	Voiced
N	n	n	n	Nasal Stop	Alveolar	Voiced
Ñ	ɲ	J	N	Nasal Stop	Palatal	Voiced
P	p	p	p	Simple Stop	Bilabial	Voiceless
PH	pʰ	p_h	ph	Aspirated Stop	Bilabial	Voiceless
P'	p'	p_>	px	Glottalized Stop	Bilabial	Voiceless
Q	q	q	q	Simple Stop	Post-velar	Voiceless
QH	qʰ	q_h	qh	Aspirated Stop	Post-velar	Voiceless
Q'	q'	q_>	qx	Glottalized Stop	Post-velar	Voiceless
R	r	r	r	Vibrant Stop	Alveolar	Voiced
S	s	s	s	Fricative Stop	Alveolar	Voiceless
SH	ʃ	S	sh	Fricative	Palatal	Voiceless
T	t	t	t	Simple Stop	Alveolar	Voiceless
TH	tʰ	t_h	th	Aspirated Stop	Alveolar	Voiceless
T'	t'	t_>	tx	Glottalized Stop	Alveolar	Voiceless
W	w	w	w	Approximant	Labial-velar	Voiced
Y	j	j	y	Approximant	Palatal	Voiced

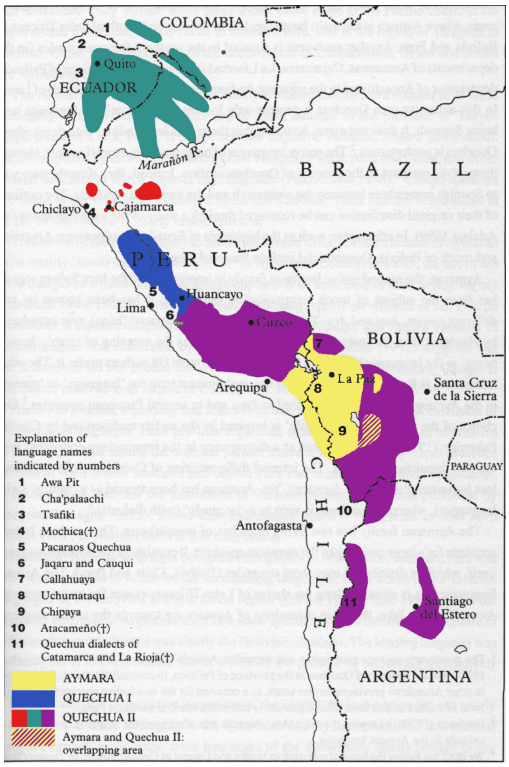


Fig. 1. Map of Quechuan Dialects/ Andean Languages (mid-20th century) (Rios 2010), adapted from Adelaar and Muysken (2004:169).

For the tokenization subprocess which is part of the text analysis, we implemented the rules for the convertion of numbers, Non-Standard words (NSW), to their standard representation, an example of a NSW is given in the table 5.

During the text analysis process, we split the text into tokens using the white space and the punctuation symbols as separators, then each NSW is identified as a separate token by token identifier rules. To identify the tokens we use the scheme regular expression in Festival, After all NSW are identify, we convert them to standard words by means of the pronunciation lexicon or (letter to sound) LTS rules.

The next step is to convert the text into a pronunciation format, for example : allin 'good', is converted to (((a) 1)((ll i n) 0)). Festival provides methods to find the pronunciations of the words. It can consist of three methods: an addenda, typically short, consisting of words added manually; a compiled lexicon, typically large; and a method for dealing with words that are not in either list called LTS Rules (Letter To Sound Rules). In our case, we use the addenda lexicon to define the pronunciation of words that are exceptions in rules. For

Table 5. Non-Standard words

NSW Category	**Written format**	**Pronunciation**
number	10	chunka
number	82	pusaq chunka iskayniyoq
number	5752	pisqa waranqa qanchis pachak pisqa chunka iskayniyoq

example, the lexicon format in festival for the word ari 'yes' is (ari n (((a) 0) ((r i) 1))).
As for the rest of the words we use the LTS rules also called Grapheme to Phoneme (G2P) rules, these rules can be specified in Festival in two ways: by providing Festival with the handcrafted rules or by building the rules automatically, for the Quechua Language we used handcrafted rules because the rules of tranformation from text format to sound format of the language is easier to implement.
The syllable structure of Quechua Language [4] is V, CV, CVC, it is implemented as a rule for a syllabification process in Festival.
For the intonation aspects, we considered that in Quechua, the stress [4] is always on the penultimate syllable, in other words, stress is not fixed, but 'shifts' with the addition of suffixes. There are a few exceptions to this general rule, mainly in connection with special emphatic utterances. These exceptions are marked by accents, e.g. with the suffix -y that expresses emphasis (Dedenbach-Salazar Senz et al. 2002:6).
Ma**ria**qa **as**wan **pi**si mi**khuq**mi **Ma**nuel **man**ta.

Maria -qa as -wan pisi mikhu -q -mi Manuel -manta
Maria -Top bit -Instr little eat -Ag -DirE Manuel -Abl

Maria eats less than Manuel.
With a simple example where one root bears different suffixes and the accent shifts: **All**qo 'dog', All**qo**cha 'little dog', Allqo**cha**lla 'only the little dog', Allqocha**lla**yki 'only your little dog'.

Recording the Voice The text corpus we used in the recording process is not phonetically balanced. For the number of existing written texts in Quechua is not enough or doesn't meet the requirements to develop this procedure. However, the text corpus we used contains all the existing phonemes that variety of Quechua used possesses. The voice for the TTS system comes from a Quechua native speaker originally from Muñani a town in southern Peru, province Azángaro in the region Puno. who has lived in the city of Cusco for 31 years. We should highlight that our speaker has a vast knowledge about the Quechua language spoken in Puno and Cusco (Quechua IIC); The recording process lasted four months approximately, for 2 hours every day in a silent room; we use the best

possible equipment that we find available. We recorded the voice of the speaker using Praat[1], a free software for acoustics analysis, after that with the help of the speaker we chose the recordings that we considered to be optimal, then using GoldWave[2], a professional digital audio editor, we proceeded to clean the recording from any noise and to minimize the volume so that the left outs of the noise disappear.
The labeling process delayed for two months as well and it was made using Praat as well, and example of this is shown in fig. 2.

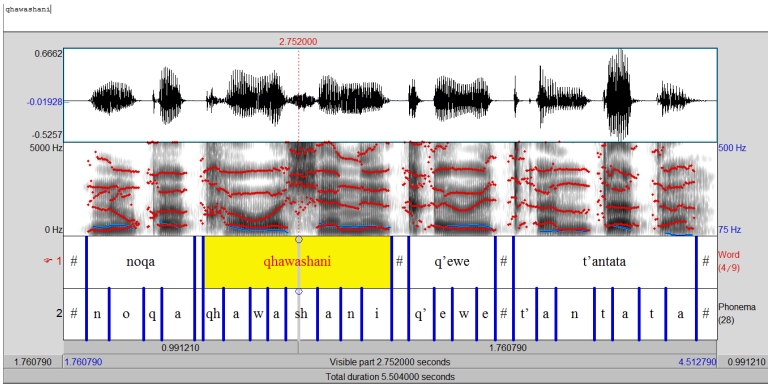


Fig. 2. Praat

Every labeled recording was tested and retested so that we have the best results possible. The emotional aspects of the new voice hasn't been implemented because Qhechua has a extremey complicated system of morphophonemics, this implies that the new voice will have an slightly Quechuan robotic accent. We selected a name for the voice using the convention, we use three parts of names, consisting of the institution name CUS, we employed the name of our city, the language QUZ following the ISO 639-6 standard for the names of languages, and the name of the speaker GUI in this case short for Guillermo.
Name of new voice in Festival: cus_quz_gui

The advantages of using the unit selection method are: first this method produces a more natural sounding speech, and second the unit selection database is created automatically using the Festival framework tools for cluster unit selection, we chose it for these two reasons, the unit selection method is described in the following section.

[1] http://www.fon.hum.uva.nl/praat/
[2] http://www.goldwave.com/

Unit Selection Database for Quechua One of the objectives of the project is to preserve a part of the spoken language in a database and give it a proper use as part of the voice corpus for the TTS system.

Cluster Unit Selection [8] This is a reimplementation of the techniques as described in [9]. The idea is to take a database of general speech and try to cluster each phone type into groups of acoustically similar units based on the (non-acoustic) information available at synthesis time, such as phonetic context, prosodic features (F0 and duration) and higher level features such as stressing, word position, and accents. The actually features used may easily be changed and experimented with as can the definition of the definition of acoustic distance between the units in a cluster.

4 Results

We get to build a working TTS System for Quechua. We have not yet performed any formal evaluation on the system, the system can only read the sentences that have all the vowels and consonants shown before in Table 3 and 4. The limitation of the use of this approach is the amount of manual work that is necessary to develop a TTS system.

5 Conclusions

The specific necessity we want to address with this work is the starting urgency for adaptation of this language to modern communication technologies and as a consequence of this shift we hope that Quechua will survive to the pressure from the dominant languages.
The project have lasted one year and a half now before showing a concrete result, the speaker who gave the voice for the TTS System is a native speaker of the language, having learned Spanish as a second language.
We had to establish a phonetic representation of the Quechua alphabet because there is not an official agreement about this topic yet, and we used this representation successfully in the implementation of the system.
We strongly recommend that researchers in order to obtain a voice that are much more natural do the following:
To increase the amount of recordings, they must be superior to 500 in number, for we have used unit selection method.
There is lots of room for improvement in the system specially in the prosodic aspects of the voice. And we hope that we could produce more TTS systems for the other varieties of Quechua.

6 Acknowledgements

This research was made as part of the undergraduate degree at the faculty of Computer Science of UNSAAC (Universidad Nacional de San Antonio Abad del Cusco), Peru. We would like to thank to Annette Rios for her invaluable insights with the Quechua language, to Magdalena Plamada, Iñaki Sainz, Antonio Bonafonte for their advice on technical issues regarding speech synthesis and the use of the Festival Speech Synthesis System. and to Guillermo Vargas Ochochoque who put the voice for this TTS System.

References

1. Campbell, Lyle: American Indian languages: the historical linguistics of Native America. pg. 253, 1997, Oxford: Oxford University Press.
2. Gordon, Raymond G., Jr.: Ethnologue: Languages of the world (15th ed.), Dallas, Texas: SIL International, 2005, http://www.ethnologue.com
3. Jaén Gomez, Alejandro: Adaptación del Sistema Texto a Voz Festival al Catalán, Departamento de Teoría de la Señal y Comunicaciones, Universidad Politécnica de Cataluña, 2007.
4. Rios, Annette: Applying Finite-State Techniques to a Native American Language: Quechua, Institut für Computer Linguistik, Universitaät Zürich., 2010.
5. Rios, Annette: Sprachtechnologie fr Quechua, Institut für Computer Linguistik, Universitaät Zürich., 2012
6. Cáceres Aragón, Armando: Runasimiq Kurkun, 2007
7. Rodriguez Champi, Albino: Quechua de Cusco, 2006
8. Unit selection databases, http://festvox.org/bsv/c2641.html, Cluster unit selection, www.festvox.org.
9. A. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis, In Eurospeech97, volume 2, pages 601–604, Rhodes, Greece, 1997.
10. Dedenbach-Salazar Sáenz, S., U. von Gleich, R. Hartmann, P. Masson, and C. Soto Ruiz: Rimaykullayki - Unterrichtsmaterialien zum Quechua Ayacuchano, 4. edition, Berlin: Dietrich Reimer Verlag GmbH, 2002.
11. Adelaar, W. F. H. and P. Muysken: *The Languages of the Andes, Cambridge Language Surveys. Cambridge University Press, 2004.

Cotovía: an Open Source Text-to-Speech System for Galician and Spanish.

Eduardo Rodríguez Banga[1], Carmen García Mateo[1], Francisco Méndez Pazó[1], Manuel González González[2], and Carmen Magariños Iglesias[2]

[1] Multimedia Technologies Group (GTM), Universidade de Vigo, Spain
[2] Centro Ramón Piñeiro para a Investigación en Humanidades, Santiago de Compostela, Spain
http://cotovia.gts.uvigo.es

Abstract. Cotovía is a text-to-speech system for Galician and Spanish developed by the University of Vigo in collaboration with the Center 'Ramón Piñeiro' for Research in Humanities. In this paper we make an official presentation of Cotovía as open-source free software under the terms of a GPL license.

1 Introduction

Cotovía is a text-to-speech system for Galician and Spanish developed by the University of Vigo in collaboration with the Center 'Ramón Piñeiro' for Research in Humanities. This collaboration stems from the interest of both institutions in promoting research in the field of speech synthesis and, particularly, in incorporating the Galician language to new technologies. Early work leading to our text-to-speech (TTS) system back to 1995 and, since then, has evolved in the framework of different public funding research projects, in which Cotovía has been used as a platform for the implementation and evaluation of new algorithms.

Although Cotovía was initially developed as an isolated application that run in its own windows environment, we soon realized the effort involved in keeping an updated version of the application running properly on different operating systems. Therefore, we decided to only continue a command line version and focus our efforts on those research lines more interesting from a scientific point of view.

Currently Cotovía is a unit-selection TTS with remarkable improvements in several algorithmic steps. Our TTS is a fairly complex program, capable of providing reasonable quality synthetic speech, as shown in the evaluations Cotovía took part (Albayzin 2008 and 2010). Of course, there are some modules that can be clearly improved, especially those that over the years were not prioritised.

Over recent years various people and institutions have contacted us expressing their interest in using Cotovía for different purposes. This is the main reason why we decided to distribute Cotovía as open source under the GPL license. This way anyone will be able to add improvements to our TTS or adapt it to its own needs. In turn, the entire community will benefit from this work.

Obviously a key issue in any unit selection TTS is the number and quality of the available speakers. At this time two Galician speakers (a male and a female) are available free of charge for any use and a Spanish male speaker is free for noncommercial applications under Creative Commons licenses.

At `http://sourceforge.net/projects/cotovia/` you can find some different Cotovía packages compiled for Debian and Windows distributions, and a Git repository with the source code and building instructions.

In the following section we discuss the main characteristics of our TTS system, making special focus on the optional techniques which are available through the command line parameters. Nevertheless, since at this time we consider to reorder the available options and possibly reduce its number, we will not show them in detail. Some preliminary description of the different options is shown when running Cotovía without any parameter as illustrated in Figure 1.

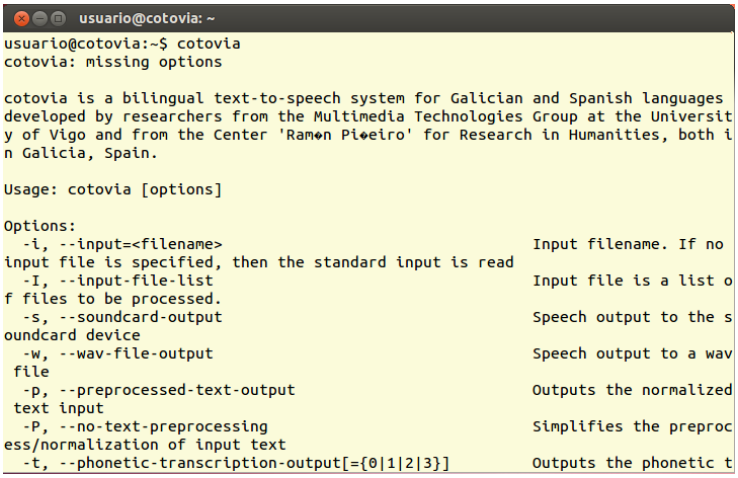


Fig. 1. Output when running Cotovía with inappropriate options

2 Cotovía Main Features

A typical concatenative TTS system consists of two main stages: linguistic-prosodic processing and acoustic processing. The first one obtains from the input text its phonetic transcription and prosody, while the second one generates the corresponding speech signal from the selected acoustic units. Basically, the structure of Cotovía also consists of these two stages, but with the peculiarity that the generation of the intonation contour is also affected by the selection of acoustic units [1] [2]. Indeed, the intonation model of our TTS, which is also based on unit selection techniques, not only generates an intonation contour for every proposition or sentence, but considers several candidate contours and selects the

final contour depending on the existence of an appropriate sequence of acoustic speech units for that contour. Thus, the choice of the contour is influenced by the segmental information of the speech. This strategy has two main advantages: on the one hand, the prosody of most of the speech segments does not need to be modified, preserving most of the quality of the original speech; on the other hand, this method introduces a great variability in the synthetic pitch contours, avoiding the generation of monotonous or easily predictable contours.

The number of candidate pitch contours considered is established when invoking Cotovía by two optional parameters (-N and -M) that determine the number of candidate contours initially generated, N, and the M candidates finally considered by excluding some contours which are too similar to each other [2] . This decimation of the candidate contours reduces the computational burden while still maintaining a high degree of variability in the intonation of the resulting synthetic speech.

We have already mentioned that the selection of acoustic units also benefits from the consideration of several intonation contours, as it is easier to find units with almost the desired prosodic features (duration and fundamental frequency, mainly) and there is no need for prosodic modification. Furthermore, as shown in [2] increasing the number of candidate contours increases the length of selected acoustic units which are actually consecutive units in the original recordings thus reducing the concatenation distortion. Moreover, although the basic acoustic unit in Cotovía is the demiphone, this procedure often results in using much longer acoustic units (composed by several original consecutive demiphones).

An apparent problem of the method described in [2] is that it seems necessary to perform sequentially a search for an optimal sequence of acoustic units for every candidate contour, thus increasing the computational load of the model. However in [3] we propose a method that performs the search of acoustic units for the different candidate contours in parallel with only a slight increment in the computational burden with regard to a single contour. Moreover, in [4] we make some considerations about the concatenation cost of the speech units that further reduce the computational complexity of the selection procedure. All these improvements are avaliable in Cotovía through the program options.

Cotovía also provides an option for controlling the prosodic modification of the acoustic units. Prosodic manipulations can be applied always, never or depending on some thresholds that take into account the difference between the desired values for fundamental frequency and duration and the real ones of the selected acoustic units.

So far we have focused on those aspects of Cotovía related to intonation modelling and synthetic signal generation, which are relatively language independent. However, much work has been also devoted to specific aspects of Galician and Spanish. Due to our interest in promoting the use of Galician in new technologies, since the beginning our work was mainly focused on this language. One of the biggest problems we faced was the lack of detailed information on some linguistic aspects of Galician (mid-vowels openness, for instance) so it was necessary to design new procedures for several steps of the linguistic processing. In

[5] we discuss the most relevant specific features of the Galician language and their implementation on our TTS. A remarkable part of the linguistic processing of Cotovía is the morphosyntactic analysis module, which includes a verbal analyser capable of handling the frequent enclitics and determining the openness of mid-vowels in the verbal stem. This module also performs grammatical category disambiguation by means of a small, but reliable, set of rules that reduces ambiguities (involving elimination of combinations that are clearly impossible), followed by a statistical model that identifies the most probable sequence of categories in a sentence. Cotovía provides an option to return the results of this analysis with different detail levels and formats. Table 1 shows a simple example of the information this module is able to provide.

Cotovía also implements the intonation model described in [6], which uses syntactic and part-of-speech features and integrates part of the phrasing algorithm in the unit selection stage.

Word	POS	Gender	Number	Lemma
A	ART_DET	FEMININO	SINGULAR	a
primeira	NUME_ORDI_DET	FEMININO	SINGULAR	primeiro
conclusión	NOME	FEMININO	SINGULAR	conclusión
da	CONTR_PREP_ART_DET	FEMININO	SINGULAR	do
análise	NOME	FEMININO	SINGULAR	análise
é	VERBO	NO_ASIGNADO	NO_ASIGNADO	ser
que	CONX_SUBOR_COMPLETI	NO_ASIGNADO	NO_ASIGNADO	que
son	VERBO	NO_ASIGNADO	NO_ASIGNADO	ser
productos	NOME	MASCULINO	PLURAL	producto
ben	ADVE_MODO	NO_ASIGNADO	NO_ASIGNADO	ben
elaborados	ADXECTIVO	MASCULINO	PLURAL	elaborado
,	COMA	NO_ASIGNADO	NO_ASIGNADO	,
que	RELA	AMBIGUO	AMBIGUO	que
respectan	VERBO	NO_ASIGNADO	NO_ASIGNADO	respectar
a	ART_DET	FEMININO	SINGULAR	a
normativa	NOME	FEMININO	SINGULAR	normativo

Table 1. Example of morphosyntactic analysis output using Cotovia (input text in Galician). Some additional information (for instance verbal mood, tense and person) is omitted in this table.

Apart from the many options available when running Cotovía, our TTS has some other yet undocumented functioning modes, that are defined and briefly commented in the file 'modos.hpp'. These modes allow, for instance, getting some statistics about acoustic units and selecting automatically a set of sentences with given structures and phonetic coverage.

The Spanish version of Cotovía, although it is not as refined as the Galician version, participated in the Albayzin Evaluations of 2008 and 2010 [7] [8] with very good results.

3 Cotovía at Sourceforge

As shown in Figures 2 and 3 users can download the data and executable files (for Debian and Windows) from the 'Files' section or clone the Git repository and follow the compilation instructions. When compiling Cotovía, an additional executable file, named 'Cotovía-ecess-m1', is also generated. This file implements a Galician 'symbolic pre-processing' module according to ECESS specifications [9] and therefore can be easily included in any TTS system that follows those specifications.

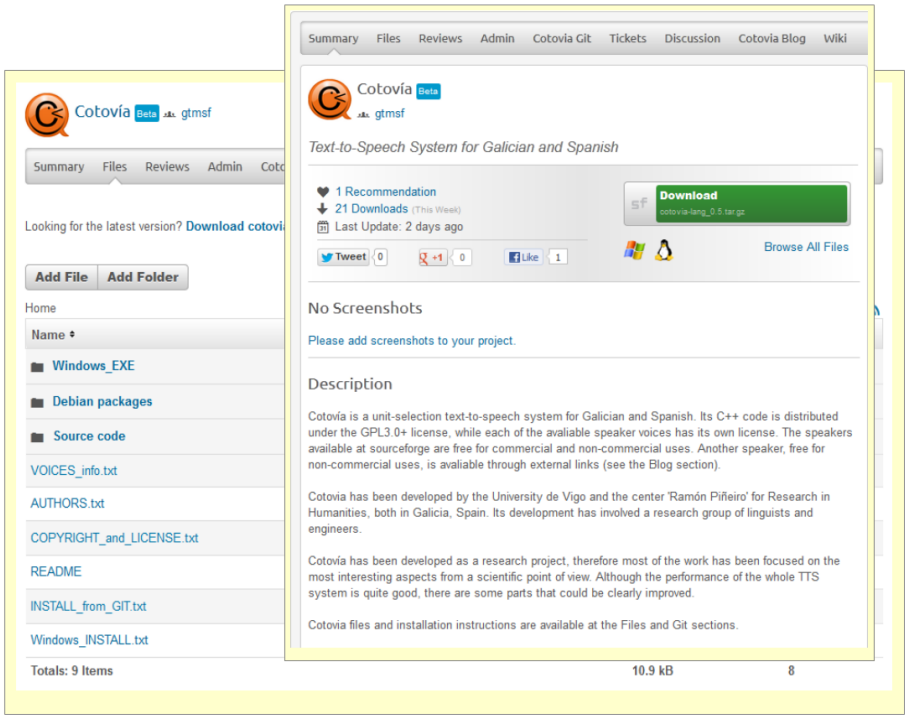


Fig. 2. Partial screen captures from Sourceforge: summary of Cotovía (foreground) and the Files section (background).

At this time two native Galician speakers (a male and a female with nicknames Iago and Sabela, respectively) are available at Sourceforge. These two speakers can be also used in Spanish although with a little quality degradation because the different language of the original corpus. Since Sabela's corpus is very large, we also provide a reduced version of this speaker. A native Spanish speaker, nickname David, is also available from Sourceforge through external links. Thus, the present available speakers are:

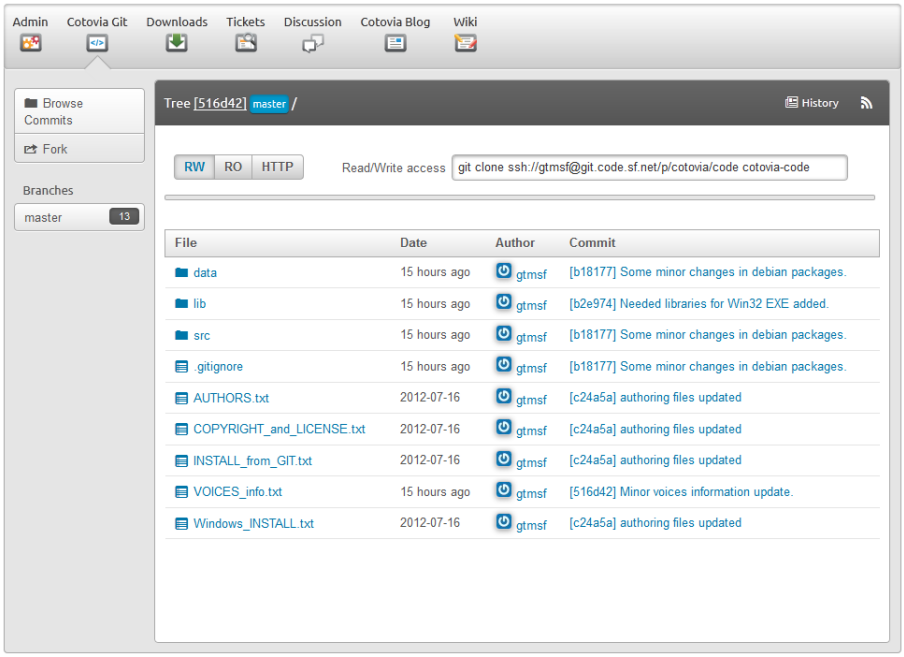


Fig. 3. Cotovia Git repository at Sourceforge

- **iago**: Galician male speaker, recording duration: 80 minutes. It was Cotovía's first voice and the quality is somewhat limited by the reduced recording time.
- **david**: Spanish male speaker, recording duration: about 2.5 hours. It was the speaker used in the Albayzin-2010 evaluation.
- **sabela-large**: Galician female speaker, recording duration: 14.5 hours. It provides good quality speech at the expense of a higher computational load.
- **sabela**: the default speaker. A subset of Sabela's recordings (about 4 hours). A good compromise between quality and computational load.

In the next future the files with the original recordings of these speakers and some other related information (text of the sentences, phonetic segmentation, pitch, etc.) will be also available at Sourceforge. This will allow their inclusion in other TTS systems and their use with any other purpose under the license terms. Nevertheless the procedure for adding new speakers to Cotovía is not yet available.

4 Integration with Speech Dispatcher

Speech Dispatcher (SD) is a client/server application for Linux that provides a high-level device independent layer to access TTS systems. It allows an easy integration of TTS systems with other applications such as Gnome-Orca, which is

mainly used by visually impaired people. The present SD distribution is prepared for several TTS engines (Festival, espeak, Flite,...) and also provides a generic driver (generic output module) to make easier to include new engines. Using this generic module we have developed our own SD version that includes Cotovía as an available speech engine. At the time of writing this paper a debian package for installation of this SD version already exists and, shortly, will be available at Sourceforge. Figure 4 shows Orca preferences with Cotovía selected as TTS engine and a few lines of the configuration file of Cotovía for SD.

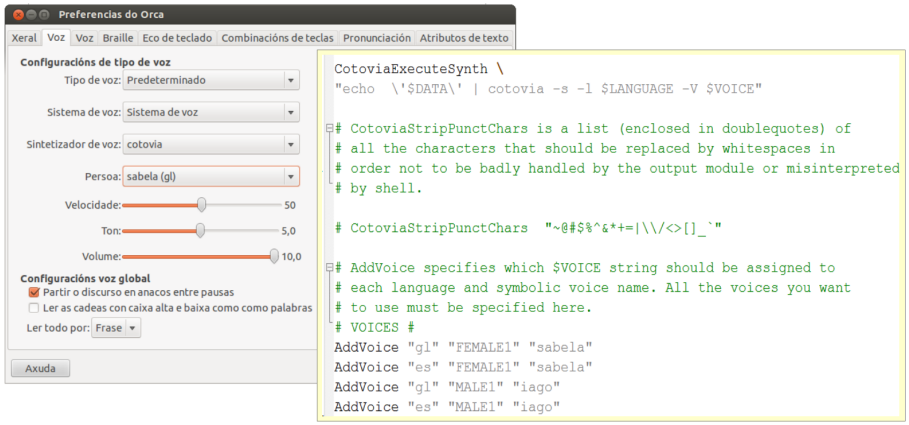


Fig. 4. Cotovia selected at Orca preferences (background) and a short portion of the configuration file of Cotovía for Speech Dispatcher (foreground).

5 Conclusions and further research

In this paper we have made an official presentation of our TTS, Cotovía, as free software under the terms of a GPL license. Three speakers are also available, each one with its particular Creative Commons license. Two native Galician speakers can be used freely for any purpose, while the native Spanish speaker is only free for noncommercial uses. In the next future we plan to continue with the development of Cotovía but we would like to have some feedback from the community to help us to establish priorities. Please, send us your comments and suggestions through the discussion forums at Sourceforge. Of course, we welcome any contribution to Cotovía development.

6 Acknowledgements

The authors wish to thank all the people that since 1995 have worked in the development of Cotovía and the funding public institutions (Spanish Government,

Xunta de Galicia and European Union by means of the European Regional Development Funds (ERDF), mainly). Present research and developments are mainly supported by the Spanish Government and ERDF under the project TEC2009-14094-C04-04 and the Center 'Ramón Piñeiro'.

References

1. F. Campillo and E. R. Banga. "Combined prosody and unit selections for corpus based text-to-speech systems". Proc. ICSLP 2002, vol. 1, pp. 141–144, Denver, 2002.
2. F. C. Diaz and E. Rodriguez Banga, "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems," Speech Communication., vol. 48, pp. 941-956, AUG, 2006.
3. F. Campillo and E. R. Banga, "Multiple f(0) contour parallel Viterbi search for unit selection speech synthesis," Electronics Letters, vol. 47, pp. 937-938, 2011.
4. F. Campillo, I. Nozhov and E. R. Banga, "Segmentwise unit selection," Electron. Lett., vol. 47, pp. 569-570, 2011.
5. M. G. Gonzalez, E. R. Banga, F. C. Diaz, F. M. Pazo, L. R. Linares and G. I. Iglesias, "Specific features of the Galician language and implications for speech technology development" Speech Communication, vol. 50, pp. 874-887, NOV-DEC, 2008.
6. F. Campillo Diaz, J. van Santen and E. Rodriguez Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information" Speech Communication, vol. 51, pp. 452-465, 2009.
7. I. Sainz, E. Navas, I. Hernáez, A. Bonafonte, and F. Campillo, "TTS Evaluation Campaign with a Common Spanish Database", ;in Proc. LREC, pp. 2155-2160, 2010.
8. F. Méndez, L. Docío, M. Arza, F. Campillo, "The Albayzín 2010 Text-to-Speech evaluation", Fala 2010 "VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop", Vigo, 2010.
9. J. Pérez, A. Bonafonte, H. Hain, E. Keller, S. Breuer, J. Tian, "ECESS Inter-Module Interface Specification for Speech Synthesis". 5th Int. Conf. on Language Resources and Evaluation 2006 , LREC 2006 . Genova. 2006

The Glissando project

Building and exploitation of a large bilingual annotated corpus for multidisciplinary prosodic analysis and applications[1]

Juan María Garrido Almiñana[1], David Escudero Mancebo[2], Lourdes Aguilar Cuevas[3]

[1]Computational Linguistics Group (GLiCom), Universitat Pompeu Fabra, Barcelona
juanmaria.garrido@upf.edu
[2]Advanced Computer Environments and MultiModal Interaction Systems group (ECA-SIMM), Universidad de Valladolid
descuder@infor.uva.es
[3]Group of Prosodic Studies (GrEP), Universitat Autònoma de Barcelona
lourdes.aguilar@uab.cat

Abstract. This paper presents the Glissando project, whose goal is the design, collection and annotation of a speech corpus for prosodic studies in Spanish and Catalan, and its joint exploitation for different research goals by the involved research groups.

Keywords. Speech corpora, Prosody, Spanish, Catalan.

1 Introduction

The aim of this paper is to introduce the Glissando project, a join initiative of three research groups interested in the analysis of Spanish and Catalan Prosody from different perspectives: the Computational Linguistics Group (GLiCom) at Pompeu Fabra University, in Barcelona; the Advanced Computer Environments and MultiModal Interaction Systems (ECA-SIMM) research group at the University of Valladolid; and the Group of Prosodic Studies (GrEp) at the Barcelona Autonomous University. The goal of the project is the design, collection and annotation of a speech corpus suitable for their research interests, as well as its joint exploitation for different research goals. This project, started in 2009, is being funded by the *Plan Nacional de I+D* of the Spanish Government (coordinated projects FFI2008-04982-C03-01 and FFI2011-29559-C02-00).

At the current state of the project, a first version of the corpus is already available for public use, including orthographic and phonetic transcription, and different types

[1] This work has been partly supported by the National R&D&I Plan of the Spanish Government (FFI2008-04982-C03-01/FILO, FFI2008-04982-C03-02/FILO and FFI2008-04982-C03-03/FILO projects).

of prosodic annotation. First exploitation results have also been obtained. Next sections describe the main features of this corpus, and outline the active exploitation research lines.

2 The Glissando corpus

Glissando is a speech corpus specially designed for the analysis of Prosody from different perspectives (Phonetics, Phonology, Discourse Analysis, Speech Technologies, comparative studies). It is actually made of two parallel corpora, **Glissando_sp** (Spanish) and **Glissando_ca** (Catalan), designed following the same criteria and structure. It is also representative of three different speaking styles: news reading, task-oriented dialogue and informal conversation. The design characteristics of the Glissando corpus a first-quality material for the experimental analysis of Spanish and Catalan Prosody:

- Good acoustic quality: Glissando has been recorded at professional studios, using high-quality microphones and recording conditions that guarantee the best quality.
- Transcribed and annotated: the Glissando corpus has been orthographically and phonetically transcribed; in addition, it includes several levels of annotation (currently prosodic unit segmentation and F0 annotation).
- High number of speakers: Glissando includes recordings of 28 different speakers per language, both professional and non-professional.
- More than 20 hours of speech have been recorded per language, an adequate size for corpus-based and experimental studies, and also for technological applications.

2.1 Contents

Both Glissando_sp and Glissando_ca are made up of three subcorpora, representing three different speaking styles: **news** (studio recordings of news readings), **task dialogues** (studio recordings of interactions between two speaking oriented to a concrete goal, the collection of information about a topic) and informal dialogues (studio recordings of informal conversations between two speakers). Table 1 summarizes the contents of all three subcorpora in both languages.

The news subcorpus was recorded using a selection of real news texts kindly provided by the *Cadena SER* radio station. These news texts were previously modified in order to obtain an adequate coverage of the target prosodic and segmental considered criteria. News texts were read in studio by eight professional speakers per language: for radio speakers and four dubbing/advertising speakers. The subcorpus includes two subsets: **'prosodic' news** (36 texts chosen according to prosodic criteria, such as the location of stressed syllables in lexical words or breath group length) and **'phonetic' news** (36 texts chosen according to segmental criteria, such as phone frequency). The design of this corpus is partially inspired in the Boston University Radio News Corpus [1].

Task dialogues subcorpus is made up of a set of interactions between two people (about 10 minutes long) oriented to the consecution of a goal: one of the speakers (*'asker'*) asks for information and the second one (*'giver'*) tries to give it to him, if he has the requested information available. It is inspired in the Map Task [2] protocol. Each pair of speakers performed three tasks: **transport** (planning a business travel between Ávila and Ciudad Real, by train or bus, for a concrete date), **university** (planning an Erasmus stay in Paris) and **tourist** (planning a tourist travel to Corfú). The subcorpus was recorded by twelve pairs of speakers (24 speakers in total): one of radio professional speakers, one of dubbing/advertising professional speakers, and 10 student pairs.

Finally, informal dialogue subcorpus is made up of informal conversations between two people with a previous knowledge of each other. The dialogue always started in the same way, remembering how both speakers first met. However, after this common starting point, conversations could derive to any other topic. This task was carried out by 6 pairs of speakers (12 people): one of radio professional speakers, one of dubbing/advertising professional speakers, and 4 student pairs. This subcorpus was designed taking as general reference the Buckeye corpus of conversational speech [3].

Corpus	**Contents**	**Speakers**	**Size**	
			Spanish	Catalan
News	36 'prosodic' texts	4 radio professional 4 dubbing/advertising professional	6 hours 40 minutes	6 hours 23 minutes
	36 'phonetic' texts	2 radio professional 2 dubbing/advertising professional		
Task-oriented dialogues	12 'transport' interactions 12 'university' interactions 12 'tourist' interactions	2 radio professional 2 dubbing/advertising professional 20 students	4 hours 37 minutes	5 hours 33 minutes
Informal dialogues	6 conversations	2 radio professional 2 dubbing/advertising professional 8 students	1 hour 6 minutes	1 hour 8 minutes

Table 1. Content summary of the Glissando corpus

2.2 Recordings

Recordings took place in two different premises: soundproof rooms at the Audiovisual Media Service of Valladolid University for the Spanish recordings, and at the Communication Campus of the Pompeu Fabra University, in Barcelona, for Catalan. In Valladolid, recordings were made on a Marantz PMD670/W1B and a Marantz PMD560 recorders, using a Mackie CR1604-VLZ mixer, at a sampling frequency of 44 KHz. In Barcelona, the Sony Vegas program running on a PC with a RME Hammerfall HDSP 9652 soundcard, and a Yamaha 02R96 mixer with ADAT MY16AT cards, were used for recordings, at a sampling frequency of 48 Khz.

All the recordings were made using two microphones for each speaker: a fixed directional one in front of them (Neumann TLM103 P48 in Valladolid; AKG C 414 B-ULS in Barcelona), and a headset wireless one (Senheisser EW100-G2, both in Barcelona and Valladolid). Headset microphones were used to ensure that the distance between the speaker's mouth and the microphone was kept constant along the recordings, making the energy registration reliable for prosodic analyses. In dialogue recordings, each speaker used different microphones, to have separate recordings of the speech from each participant, so as to avoid the overlapping of signals. A laringograph (Laryngograph Processor, from Laryngograph Ltd) was also used to record the glottal activity in some of the news recordings (those of the category B speakers). This signal can be used to detect the glottal closure instants and to get an accurate pitch estimation. In total, four synchronous channels (six if the laryngograph was included) were recorded.

Recordings were stored on wav files, one per signal (one wav for the fixed microphone, one for the headset microphone and one for the laryngograph, if any). In the case of dialogue recordings, stereo wav files were created, including the signal of each speaker's microphone. Then, two stereo wav files were obtained for each dialogue, one for the fixed microphones and one for the headset microphones.

2.3 Speakers

As already mentioned, 28 different speakers participated in the recordings of each language subcorpus, 8 professional (4 radio speakers and 4 dubbing/advertising speakers) and 20 undergraduate students. Radio speakers came from the *Cadena SER* radio station in Valladolid, in the case of Spanish, and from different radio stations in Barcelona (*Catalunya Ràdio*, *RAC1*, *Ràdio Estel*) in the case of Catalan. Dubbing/advertising speakers were all active professional in the field in Valladolid, in the case of Spanish, and Barcelona, in the case of Catalan speakers. The most general profile of the student speakers was last-year undergraduate students of Communication (at *Universidad de Valladolid* in the case of Spanish; at *Universitat Pompeu Fabra* and *Universitat Autònoma de Barcelona* in the case of Catalan), although some other student profiles were also accepted in some cases.

The selected speakers were all native speakers of the target language, speakers of the Castilian variety in the case of Spanish, and of Central Catalan in the case of Catalan.

All the speakers were organized into four categories, defining their contribution to the corpus recording. Table 2 describes these categories. According to this organization, some speakers (those of the A category, all of them professional) participated in the recordings of al three subcorpora; some others in recorded two subcorpora (those of the C category, all of them students); and finally, the last two groups participated only in the recordings of one subcorpus (the news subcorpus, in the case of the B category, and the task oriented dialogues, in the case of D category). By this distribution, we ensured speaker comparability among all three speaking styles, at least for the professional profiles.

2.4 Orthographic transcription

In addition to the recordings, the corpus includes also the corresponding orthographic transcriptions of the collected materials. In the case of the news subcorpus, the original texts read by the speakers were revised in order to adapt their contents to the actual reading of each speaker. The resulting transcription is stored in plain text files. Dialogues were all transcribed manually, applying the criteria proposed by TEI [4]. The obtained transcription was stored in xml files, which include, in addition to the transcription itself, additional information about the organisation of the dialogues into speech turns, and the annotation of the paralinguistic events.

2.5 Annotation

The annotation of the corpus is stored in TextGrid Praat files [5]. Five levels of annotation are currently available, stored in separate tiers of the TextGrid file and time-aligned with the speech signal: orthographic transcription (word by word), phonetic transcription (using the SAMPA phonetic alphabets for Spanish [6] and Catalan [7]), syllable boundaries, intonation group boundaries, and breath group boundaries.

A first version of this annotation was obtained automatically, by using several transcription, annotation and alignment automatic tools. Orthographic and phonetic transcriptions were obtained from the output of the phonetic segmentation tool embedded in the Cereproc's Voice Creation Kit. The versions of this tool for Spanish and Catalan were developed jointly by Cereproc and the Speech and Language Group of *Fundació Barcelona Media*, with the participation of members of GLiCom [8]. Prosodic unit segmentation (syllables, intonation groups, breath groups) was obtained using SegProso, an automatic annotation tool developed at GLiCom.

Category	Tasks	Speaker type	Number
A	News (prosodic) Informal dialogue Task-oriented dialogue	Professional (radio) Professional (dubbing/advertising)	2 2
B	News (prosodic+ phonetic)	Professional (radio) Professional (dubbing/advertising)	2 2
C	Informal dialogue Task-oriented dialogue	Student	8
D	Task-oriented dialogue	Student	12
Total			28

Table 2. Speaker categories for each language

2.6 Public version

A first version of the corpus has been made accessible for public access at the web page of the project (http://veus.barcelonamedia.org/glissando/), under a Creative Commons license, which allows its free use, modification and distribution if the original source is mentioned. Previous registration is required to access the corpus. This public version includes all the collected recordings, their orthographic transcription and the 'automatic' version of the annotation files. Tasks-oriented and informal dialogues are available in two versions: **'Complete'**, in which each dialogue recording is included into a single file; and **'Turns'**, which includes separate wav files for every speaking turn in a dialogue. Table 3 describes the available contents for each corpus unit in the different subcorpora.

2.7 Ongoing improvements

There are currently two ongoing tasks in the project oriented to improve the current version of the Glissando corpus: the manual revision of the transcription and annotation of the corpus obtained by automatic means, and the addition of new levels of annotation.

The revision task is aimed to detect and fix possible errors in the transcription and annotation generated by the automatic tools. All the tiers included in the time-aligned

annotation TextGrid (word, phone, syllable, intonation group and breath group boundaries) are being revised by linguists specially trained for this task. Most frequent errors found in the already corrected material include the insertion of wrong pause boundaries, mismatches between the theoretical word and phonetic transcription (the ones derived from the orthographic transcription) and actual realisations of the speakers in order to find and correct, and false detections of intonation group boundaries. At the current stage of the work, all the news subcorpus and a small part of the dialogues subcorpora of Glissando_sp (Spanish), and a subset of the news subcorpus (5 speakers out of 8) in the case of Glissando_ca, have been already revised.

Special attention is also being paid to the addition of new annotation tiers to the corpus. One of the main goals established at the beginning of the project was to include in Glissando annotation several types of intonation standards, including the most used ones, such as ToBI [9] or MoMel [10], in order to offer researchers working on different prosodic analysis frameworks a corpus suitable for their research interests, and also to attempt comparisons between some of these systems. Intense research on the automatic annotation of corpora using ToBI labels has been carried out in parallel to the development of the corpus [11] [12] [13] [14]. This research has led to a first automatic annotation of prominences using the ToBi-framework conventions, and it is planned to apply these results to the full automatic ToBI annotation of the Glissando corpus. Some efforts have also been done in the definition of the ToBI label set and labelling conventions. In addition, the corpus has been partially annotated using MelAn, an automatic tool for the annotation of intonation inspired in the IPO model [15]. This annotation allows to keep raw F0 values corresponding to the relevant inflection points in the F0 contours, and their annotation in terms of 'peaks' (P) and 'valleys' (V).

Finally, some advances have been done in the task of including other types of labelling to the corpus, not directly related to prosodic annotation, but relevant for prosodic analyses. This is the case, for example, of speech act labelling: a pilot annotation test, using a first list of speech act labels, has been done with a sample dialogue in Spanish; also, a pilot version of a tool for the automatic for detection of speech acts in text has been developed.

Future tasks to be started in the near future in this area include the annotation of the corpus with morphosyntactic information, and the annotation of focalised elements, including the adaptation or development of automatic tools to support these annotations.

3 Corpus exploitation

As already mentioned, the Glissando corpus was designed to fulfill the research interest of the groups involved in its development, different but with common points:

- Modelling of Spanish and Catalan Prosody, based on the analysis of large amount of data (phonetic modelling; phonological modelling, within the ToBI framework; analysis of the relation between phonetic and phonological patterns).
- Experimental comparison of prosodic annotation systems

- Development of prosodic models for Text-to-Speech applications
- Experimental analysis of the prosodic features of the dialogue speech in Spanish and Catalan
- Experimental analysis of the prosodic features of news readings by professional speakers
- Inter-linguistic comparison of prosodic features
- Development of teaching materials

Exploitation tasks are in a very initial phase: some work has been carried out in the area of phonetic modeling of F0 contours in Spanish dialogues, in the development of prosodic models for Text-to-Speech, and in the description of phonetic features of Prosody in news reading in Spanish and Catalan. A preliminary description of the F0 patterns used by the Spanish professional speakers in the task-oriented and informal dialogues is given in [16], and first results about the F0 patterns used by professional speakers in news readings in Catalan have been recently described in [17]. Also, some work has been done in the description of the prosodic features of news material by professional speakers, and in the description of the paralinguistic elements used in dialogues, although the obtained results, still preliminary, have not been published yet. And new research tasks, related to the different research areas detailed before, are expected to be started in the next future.

4 Conclusions

To date, the Glissando project has already offered some useful results, mainly in the form of a high-quality speech annotated material that is starting to be exploited by the involved research groups. However, this corpus may also be of interest for other research groups interested in the description of Spanish and Catalan Prosody from different perspectives.

These improvements will be made public in future releases of the corpus, available from the project website.

					News	Dialogues ('complete' version)	Dialogues ('turns version)
Unit					News text	Dialogue	Speech turn
Contents	Speech signal	Format wav (16.000 Kz)	Fixed microphone signal	.fix.wav	X (Mono)	X (Stereo)	X (Mono)
			Headset microphone signal	.wir.wav	X (Mono)	X (Stereo)	X (Mono)
			Laryngograph signal	.lar.wav	X (Mono)		
	Orthographic transcription	Text format (UTF-8)	Text	.txt	X		
		XML format (UTF-8)	Text enriched with information about speech turns and paralinguisitc events	.xml		X	
	Time-aligned annotation	TextGrid Praat format	Tiers containing different levels of annotation	.TextGrid	X	X	X
	Intensity values	Intensity Praat format	Intensity values corresponding the the fixed microphone signal	.fix.Intensity	X	X	X
			Intensity values corresponding the the headset microphone signal	.wir.Intensity	X	X	X
	F0 values	Fixer Pitch de Praat	F0 values corresponding the the fixed microphone signal	fix.Pitch	X	X	X
			F0 values corresponding the the headset microphone signal	.wir.Pitch	X	X	X

Table 3. Units, contents and files available in the public version of each language subcorpus

5 References

1. Ostendorf, M., Price, P., Shattuck, S.: The Boston University Radio News Corpus. Tech. rep., Boston University (1995)
2. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weinert, R.: The hrc map task corpus. Language and Speech (24), pp. 351 – 366 (1991)
3. Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye corpus of conversational speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, oh: Department of psychology, Ohio state University (distributor) (2007)
4. Sperber-McQueen, C, Burnard, L.: Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative (1994)
5. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program] http://www.praat.org/ (2012)
6. http://www.phon.ucl.ac.uk/home/sampa/spanish.htm

7. http://liceu.uab.es/ joaquim/language resources/SAMPA Catalan.html
8. Garrido, J. M., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M., Pidcock. Ch.: The CEREVOICE speech synthesiser. In: Actas de las V Jornadas de Tecnología del Habla (Bilbao, 12-14 noviembre 2008), pp. 126-129 (2008)
9. Beckman, M,, Hirschberg, J., Shattuck-Hufnagel, S.: The original ToBI system and the evolution of the ToBI framework. In Jun SA (ed): Prosodic Typology: The Phonology of Intonation and Phrasing, New York: Oxford University Press, pp 9–54 (2005)
10. Hirst, D., Espesser, R.: Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix, 15, pp. 71-85 (1993)
11. Escudero, D., Aguilar, L., Vanrell, M., Prieto, P.: Analysis of inter-transcriber consistency in the Cat ToBI prosodic labeling system. Speech Communication, 54, pp. 566–582 (2012)
12. González, C., Escudero, D., Vivaracho, C., Cardeñoso, V.: Improving automatic classification of prosodic events by pairwise coupling. IEEE Transaction on Audio, Speech and Language Processing p (in press)
13. Escudero, D., Aguilar, L., Ferreras, C.G., Vivaracho, C., Cardeñoso, V.: Cross-lingual English Spanish tonal accent labeling using decision trees and neural networks. In: Travieso-González C.M., Hernández, J.B.A. (eds): Advances in Nonlinear Speech Processing - 5th International Conference on Nonlinear Speech Processing, NOLISP 2011, Las Palmas de Gran Canaria, Spain, November 7-9, 2011. Proceedings, Springer, Lecture Notes in Computer Science, 7015, pp 63–70 (2011)
14. Escudero, D., Vivaracho, C., González, C., Cardeñoso, V., Aguilar, L.: Analysis of inconsistencies in cross-lingual automatic ToBI tonal accent labeling. In: Habernal, I,, Matousek, V. (eds): Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings, Springer, Lecture Notes in Computer Science, 6836, pp 41–48 (2011)
15. Garrido, J. M.: A tool for automatic f0 stylisation, annotation and modelling of large corpora. In: Speech Prosody 2010, Chicago (2010)
16. Garrido, J. M., Rustullet, S.: Patrones melódicos en el habla de diálogo en español: un primer análisis del corpus Glissando. Oralia: Análisis del discurso oral, 14, pp. 129-160 (2011)
17. Garrido, J. M.: GLISSANDO. Un corpus anotat per a l'anàlisi de la prosòdia del català i del castellà. Descripció i primers resultats d'explotació. Workshop d'entonació del català, juny 2012 (2012)

The CASMACAT Project: The Next Generation Translator's Workbench

Daniel Ortiz-Martínez, Germán Sanchís, Francisco Casacuberta,
Vicent Alabau, Enrique Vidal, José-Miguel Benedí, Jesús González-Rubio,
Alberto Sanchís, and Jorge González

D. Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de vera s/n, 46022, Valencia, Spain
{dortiz,gsanchis,fcn,valabau,evidal,
jbenedi,jegonzalez,josanna,jgonzalez}@dsic.upv.es

Abstract. The goal of the CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation) research project is to build the next generation translator's workbench to improve productivity, quality, and work practices in the translation industry. The CASMACAT project is co-funded by the European Union under the Seventh Framework Programme (FP7) Project and involves a total of four partners, namely, the University of Edinburgh (United Kingdom), the Polytechnic University of Valencia (Spain), the Copenhagen Business School (Denmark) and the Spanish translation company Celer Soluciones. The work to be developed during the project is distributed in eight workpackages and ranges from user interface studies and user modelling to new machine translation techniques that allow human users and machine translation systems to collaborate in order to obtain high quality translations.

1 Introduction

European integration and globalisation beyond it increases cross-border commercial, cultural, and political interaction. However, while the significance of political borders diminishes, the risk remains that the world will stay fractured by linguistic boundaries. The need to address each individual in a language that she speaks, and ideally in her native language, requires a huge amount of translation work.

While there have been significant improvements to machine translation technology, the vast majority of this work is targeted towards bulk translation that is good enough or fit for use. A user on the Internet is satisfied with a rough translation, if it fills her information need. Opposed to that is the demand for high quality translations by the marketplace: the translation of reports and announcements of multi-national organizations, marketing material and product descriptions of commercial companies, and many other localization needs. Such high quality translations are still almost exclusively provided by human translators.

Productivity of human translators can be increased with computer aided translation (CAT) tools: translation memories are standard in the translation industry, but post-editing machine translation output is only slowly becoming an increasingly used practice [10]. The current integration of machine translation technology into human translators work processes is often done overly simplistic, breaks their work practices, and it is widely resisted. Hence, the CASMACAT project will carry out in-depth study of translator behaviour to tailor the tools to the requirements of translators, and not the other way around. In the CASMACAT project a novel workbench that will increase the productivity of human translators is proposed by addressing their needs for the right type of assistance at the right time.

2 Project Goals

In recent years, there has been measurable progress in the field of machine translation, both in terms of quality and increased use, due to the widespread adoption of statistical methods. Vigorous research is carried out in academic and commercial research labs. As it will be described below, there has been some progress in aiding human translators, but the vast potential of creating a new workbench for human translators is mostly unfulfilled. The CASMACAT project[1] will try to generate such workbench by transferring the methods from the statistical machine translation community to the task of assisting human translators. Whereas the translation technology is ripe enough, design issues of the user interface and its acceptance by the translator have been widely neglected. The development of such tools must not simply follow technical possibilities, but it should be driven by a better understanding of the behaviour of human translators.

This is the goal of CASMACAT project: to carry out cognitive analysis that provides insight into the human translation process to guide our development of a new workbench for translators. The partners of the consortium are drawn from the leading groups in cognitive modelling of translators, statistical machine translation and computer aided translation. The commercial partner Celer Soluciones, a translation agency with significant experience in innovative computer aided translation tools, will evaluate the novel methods to provide feedback and guidance.

Currently, Europe is leading in providing translation services. To maintain and extend this leadership role, to a large degree driven by the linguistic diversity in a common economic and political union, European translation companies must continue to improve their work practices and adopt advanced technology. The CASMACAT projects aims to contribute to this process by developing novel types of assistance and bring them to the users.

2.1 Scope

Human translation is performed by different types of translators, tackles different text types, and deals with different language pairs.

[1] http://www.casmacat.eu/

- **Translators**: The needs of user communities will be addressed, ranging from professional translators to volunteer translators.
- **Text Types**: Much of what professional translators translate is repetitive, technical material. In contrast, volunteer translators are more commonly interested in generally accessible material, or technical content in their own area of expertise.
- **Language Pairs**: The methods developed in the CASMACAT project are generally language-independent. Work will be done on languages for which the necessary data resources are available and for which project participants have sufficient expertise.

2.2 Workbench

The CASMACAT project will develop a new open source workbench for human translators. All functionalities developed by the project will be integrated in a web-based online service which may also be installed locally on the desktop of a translator. The availability as web service will make it easy to integrate it into existing translation workflows. The academic partners of the consortium have previously developed such tools and will synthesize their experience.

Translog [5, 6, 2] is the leading tool for analyzing text production processes developed by CBS, which has been used in a large number of translation process studies [11]. MIPRCV-IMT is a reference implementation of novel techniques in interactive-predictive machine translation (or interactive translation prediction from now on) that were developed by UPVLC [1, 9, 8]. Caitra [7] was developed by UEDIN in order to explore new types of assistance for human translators.

The new CASMACAT workbench will integrate all relevant functionalities of the pre-existing tools. All new features explored in the project will be implemented within the new system.

2.3 Cognitive Analysis

An important objective of the CASMACAT project is to gain insight into the cognitive processes involved in human translation. How large are the text segments actively considered by a translator (the whole text, individual sentences, or only subsentential segments of limited length)? What are the subtask that a translator spends most time on (e.g., understanding the source text, looking up unknown words, investigating lexical translations, syntactic restructuring of the sentence, ensuring fluency of the output)? How does translation differ from well-studied simpler cognitive processes such as reading and text production?

The cognitive analysis will inform the design of the CASMACAT translation workbench in a range of ways. It will determine what types of assistance are offered to the translator, what information should be displayed on the screen, and what information should be hidden as it would be distracting. The different versions of the user interface will be evaluated in user studies using eye-tracking and other commonly used methods in cognitive analysis.

2.4 Advanced Computer Aided Translation

CASMACAT will use well-established statistical methods and explore novel approaches in order to generate and disambiguate translation proposals. Dynamically generated translation options will be sent to and visualized in an interactive translation assistance tool. Input from the user will be given by means of keyboard, mouse or electronic pen (e-pen). Two basically different approaches to computer assisted translation (CAT), Interactive Translation Prediction and Interactive Editing, will be developed, compared and evaluated and a cognitive model of the translator will be developed to predict the translators performance.

A novel reworking of the idea of interactive translation prediction (ITP) will allow for the construction of systems that produce high-quality results by placing a human operator at the centre of the production process. The ITP paradigm embeds a statistical MT engine within an interactive editing environment. The human serves as the guarantor of high quality; the role of the automated systems is to ensure increased productivity by proposing well-formed extensions to the current target text, which the operator may then accept, correct or ignore. Interactivity allows the system to take advantage of the human-validated portion of the text to improve the accuracy of subsequent predictions.

3 Project Partners

The project is made up of 4 partners from 3 European countries: 3 universities (United Kingdom, Denmark, Spain) and 1 company (Spain):

1. University of Edinburgh, United Kingdom
2. Polytechnic University of Valencia, Spain
3. Copenhagen Business School, Denmark
4. Celer Soluciones, Spain

The partners are complementary due to their different scientific expertise statistical machine translation, computer aided translation, open source software development, cognitive studies of natural language processing, translation process research, and delivery of translation services. Each partner's expertise and their research interest are a good fit with their assigned tasks.

– **Academic Partners**
  1. **UEDIN**: University of Edinburgh, School of Informatics[2]
     Drs. Philipp Koehn, Frank Keller
     Areas of Expertise: statistical machine translation, open source development, machine learning, computer aided translation, evaluation, cognitive modeling of language processing.
  2. **UPVLC**: Universitat Politécnica de Valéncia, Technical Institute for Informatics[3]

[2] http://www.inf.ed.ac.uk/
[3] http://www.iti.upv.es/

Drs. Francisco Casacuberta, Enrique Vidal
Areas of Expertise: statistical machine translation, interactive machine translation, finite state approaches to machine translation, handwritten text recognition.

3. **CBS**: Copenhagen Business School, Department of International Language Studies and Computational Linguistics[4].
   Drs. Michael Carl, Arnt Lykke Jakobsen, Jakob Elming, Christopher Teplovs
   Areas of Expertise: translation process research, statistical machine translation, data visualisation, web-design and collaborative web applications.

– **Company**
1. **CS**: Celer Soluciones[5]
   Roberto Silva, Enrique Diaz de Liaño
   Areas of Expertise: translation services, evaluation and use of novel computer aided translation techniques.

4 The CASMACAT Workbench

The CASMACAT Workbench allows users to enter documents in a source language, and then receive assistance in translating them into a target language. The assistance is based on information from machine translation systems.

Figure 1 shows a diagram of the architecture of the CASMACAT Workbench. According to the figure, there are five major components:

– **Editor**: is the interface between the user and the functionalities provided by the CASMACAT workbench. It communicates with two different components: the GUI server and the translation server. Input can be given via keyboard, mouse or e-pen. In addition to this, the behaviour of the translator can be studied by means of an eye-tracking system.
– **GUI server**: serves web pages to the editor interface. Handles information useful for cognitive analysis of the translation process, including logging and replay information. Such information is stored into the workbench databases.
– **Translation server**: provides translation services including regular MT and interactive MT by means of a specific application programming interface (API). In addition to this, it is also used for authentication and document management purposes (document preprocessing, storing of translation information per each source document, etc.).
– **HTR server**: handles user interactions by means of an e-pen. It offers a specific API to the CASMACAT editor.
– **Database**: two databases have been identified. The first one should be visible to the translation server and store user information, documents, partial or total translations of the documents. The second one should be visible to the GUI server and store replay information.

[4] http://uk.cbs.dk/research/departments_centres/institutter/isv/
[5] http://www.celersol.com/?idioma=en

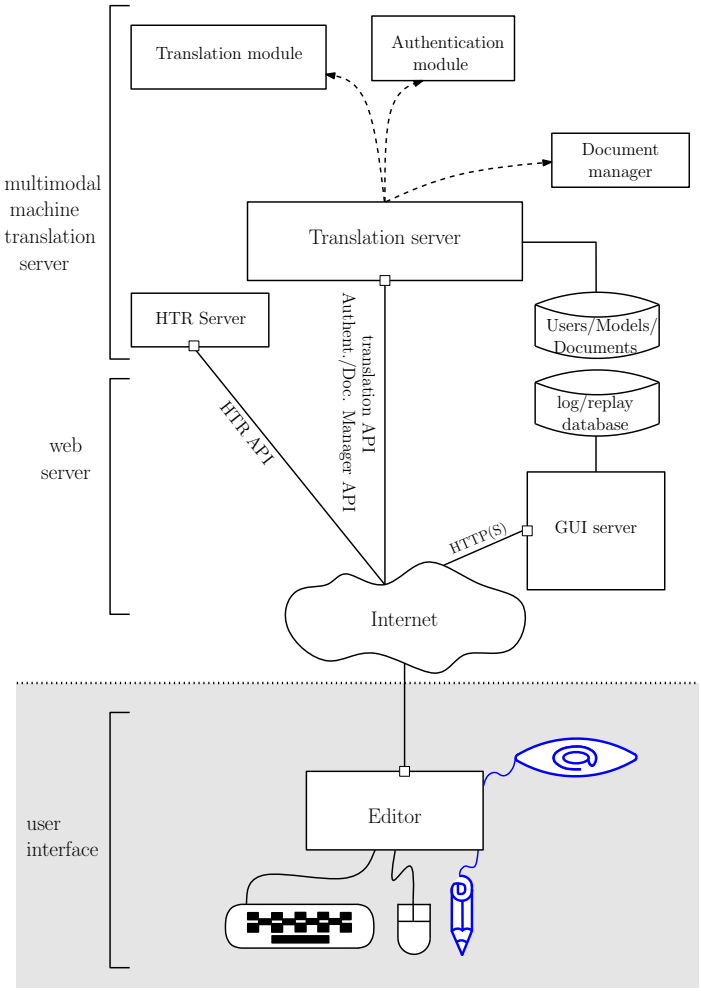


Fig. 1. CASMACAT workbench architecture.

5 Work Packages

The CASMACAT project is structured in a total of eight work packages. In the rest of this section we briefly describe each work package and provide some information about the timing of tasks.

5.1 Work package Description

WP1: User Interface Studies, Cognitive and User Modeling
This WP lays the empirical foundations for the development of the CASMACAT

workbench. A series of experiments will establish basic facts about translator behaviour in computer-aided translation, investigating the usefulness of visualisation option in post-editing and interactive translation, for different types of text and for translators with different degrees of expertise.

WP2: Interactive Translation Prediction
This work-package is devoted to investigating innovative approaches to human-MT interaction covering three important dimensions: effectiveness, efficiency and communication. Improved interaction protocols and more efficient search procedures will allow the systems to supply more useful and faster predictions. Finally, sophisticated interaction modalities will enrich the information supplied to the system and will facilitate the generation of user feedback, thereby enhancing the overall performance and usability. Some of the topics that will be studied withing WP2 include alternative search and machine learning criteria from prediction, use of e-pen as additional input modality, use of new SMT models for ITP, etc.

WP3: Interactive Editing
In this work package, three main lines of research are pursued: (1) enriched post-editing, where augmented information about the machine translation output aids the post-editor to highlight possible weaknesses of the translation, (2) authoring assistance, where knowledge about the document, prior translations, and general knowledge about language use helps the editor to find better translations, and (3) automatic reviewing, where a second pass over the translation is aided by automatically detecting untranslated or added material, as well as ensuring consistent translation of terminology.

WP4: Adaptive Translation Models
Human interaction offers another unique opportunity to improve the performance of the ITP systems by tuning the translation models. This WP tackles the problem of how text validated by the human translator can be used to adapt the system to changing environment.

Techniques based on on-line learning, cache-based approaches, Bayesian adaptation and/or reinforcement learning will be explored for their full integration with the statistical translation model training.

WP5: Integration
All the methods developed by the CASMACAT project will be integrated into a new translators workbench, the CASMACAT workbench. Its development will take place throughout the project in three distinct stages: specification of requirements, implementation of core functionality, and integration of novel methods.

WP6: Evaluation
In this work package, the CASMACAT workbench will be exposed to a wider

community of users and engage the localization industry to gain wider adoption. In addition to this, field trials will carried out to study the use of the workbench in a real-world environment, promote it to community translation platforms and the language service industry.

WP7: Dissemination
This work package is devoted to explore ways to reach the different type of potential users of the CASMACAT workbench as well as to publish scientific results.

WP8: Management
This work package is devoted to the different aspects of project management.

5.2 Timing of Tasks

CASMACAT is a 3-year project from November 2011 to October 2014. The workplan is strongly focused on the development of the CASMACAT workbench (WP5). Specifically, three different releases will be developed:

1. **Initial release** (month 6): restricted to the capability of post-editing machine translation output provided by a static system and logging of translator activity.
2. **Beta release** (month 18): which integrates advanced interactive machine translation methods and a number of different features, including the use of confidence measures, integration with translation memory, online learning for interactive prediction, support for eye tracking, etc.
3. **Final release** (month 30): which integrates all remaining advances of the CASMACAT project.

6 Related Projects

The CASMACAT project is funded in the same cycle with two related projects which also broadly aim at improving the interaction of machine translation technology and human translators. The MATECAT (Machine Translation Enhanced Computer Assisted Translation) project[6] mainly aims at the improvement of statistical machine translation technology to more closely suit the needs of translators working in a post-editing workflow. The ACCEPT (Automated Community Content Editing Portal) project[7] aims at improving the role of machine translation in community and volunteer translation platforms.

While the specific aims of the three projects are quite different, there are a few tasks where they pursue similar goals and where co-ordination between the three projects is essential to obtain synergy and not duplication of efforts. It

[6] http://www.matecat.com/matecat/the-project/
[7] http://www.accept.unige.ch/index.html

is expected a especially close co-ordination between the MATECAT and CASMACAT project, including joint project meetings and joint dissemination events.

In addition to the above mentioned MATECAT and ACCEPT research projects, there is also another project which is not completely focused on machine translation but is also related to the CASMACAT project, the TRANSLECTURES (Transcription and Translation of Video Lectures) project[8]. The aim of TRANSLECTURES is to develop innovative, cost-effective solutions to produce accurate transcriptions and translations in the VideoLectures.NET web portal, a free and open access educational video lectures repository.

References

1. Alabau, V., Ortiz-Martínez, D., Sanchis, A., Casacuberta, F.: Multimodal interactive machine translation. In: ICMI-MLMI '10: Proceedings of the 2010 International Conference on Multimodal Interfaces. Beijing, China (2010)
2. Carl, M., Jakobsen, A.L.: Towards statistical modelling of translators activity data. International Journal of Speech Technology 12(4) (2010), http://www.springerlink.com/content/3745875x22883306/
3. Dam, H.V., Engberg, J., Gerzymisch-Arbogast, H. (eds.): Knowledge Systems and Translation, Text, Translation, and Computational Processing, vol. 7. Mouton de Gruyter, Berlin, New York (2005)
4. Hansen, G. (ed.): Probing the process in translation: methods and results, Copenhagen Studies in Language, vol. 24. Copenhagen: Samfundslitteratur (1999)
5. Jakobsen, A.L.: Logging target text production with Translog. In: [4]. pp. 9–20 (1999)
6. Jakobsen, A.L.: Investigating expert translators processing knowledge. In: [3]. pp. 173–189 (2005)
7. Koehn, P.: A web-based interactive computer aided translation tool. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations. pp. 17–20. Association for Computational Linguistics, Suntec, Singapore (August 2009), http://www.aclweb.org/anthology/P/P09/P09-4005
8. Lagarda, A.L., Civera, J., Juan, A., Casacuberta, F.: Interactive pattern recognition and human language technology for di gital audiovisual content processing. Journal of Machine Learning Research 11: Workshop on Applications of Pattern Analysis, 103–110 (09 2010)
9. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Online learning for interactive statistical machine translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT). pp. 546–554. Los Angeles (Jun2 2–4 2010)
10. Plitt, M., Masselot, F.: A productivity test of statistical machine translation post-editing in a typical localization context. The Prague Bulletin of Mathematical Linguistics 93, 7–16 (2010)
11. Schou, L., Dragsted, B., Carl, M.: Ten years of Translog. Copenhagen Studies in Language 38, 37–48 (2009)

[8] http://llach.dsic.upv.es/~translectures/

Google's Voice Search in Iberian Languages

Ignacio Lopez-Moreno, Fadi Biadsy, Eugene Weinstein, and Pedro J. Moreno

Google Inc.
New York, NY, USA
{elnota,biadsy,weinstein,pedro}@google.com

Abstract. This paper introduces a large scale effort to develop a set of Automatic Speech Recognition Systems for all Iberian Languages. Our goal is to provide Large Vocabulary ASR systems that support Google Voice Search for the five Iberian languages: Basque (`Eu`), Catalan (`Ca`), Galician (`Gl`), Portuguese (`Pt`) and Spanish (`Es`). All together, these systems have the potential to reach 55 million people. We use anonymous Google search queries to build the language models, whereas we collect and annotate our own acoustic data to train the acoustic models. For lexicons, we utilize pronunciation rules as well as hand-crafted pronunciation dictionaries. Since the Iberian Spanish system has been in production for a number of years we use a different approach to build and update its acoustic and language model. In this case we use unsupervised data selected from the production logs. Our experiments shows good performance for all five systems, even in the bi-lingual environments of Spain. Finally, we introduce the embedded `Es` system, which unlike the standard `Es` system, runs entirely in a mobile phone. We obtain an average word error rate of 20.34.

Keywords: Speech Recognition, Basque, Catalan, Galician, Portuguese and Spanish, Voice Search

1 Introduction

Automatic voice-enabled services have been the subject of research for several years. They have resulted in various commercial services, including Goog 411 [1], the first fully automated service to locate businesses, and, subsequently, Voice Search [2], a popular service enabling users to perform web searches simply by speaking them to a mobile phone. After Voice Search was released in the English language in 2008, the natural next step was its internationalization.

Our end goal is to support the official languages of all countries; however, the way languages are prioritized depends on a number of factors, including population, the number of mobile phones per capita and the availability of training and testing data. Currently, Voice Search services an extraordinary number of queries in 50 different *locales* (i.e. a language & country code pair); however, Spanish (`Es`) is the only Iberian language included in this number. As an illustration, Figure 1 depicts the growth of our initial `Es` system's live traffic during

the last 16 months. In general, as the number of language-specific systems grows, so does the user acceptance of the speech-related services and thus the system's overall performance.

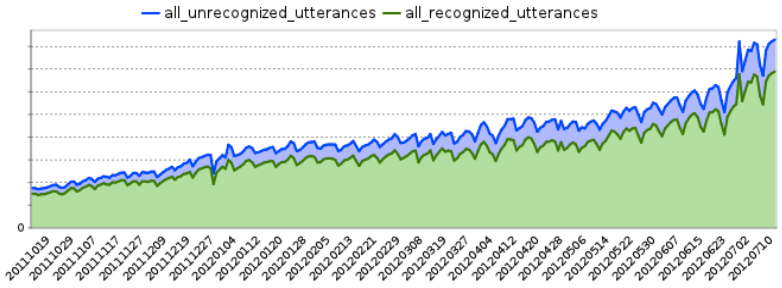


Fig. 1. Relative live traffic data volume for `Es` system (y axis) from Nov. 2011 to Jul. 2012 (x axis). In green, we show the relative number of queries per day. In blue, we add the number of queries rejected by our confidence model, mostly accounting for noises, silences or indistinguishable speech.

In this paper, we describe the process of building completely new language-specific systems for the following four Iberian languages: Basque (`Eu`), Catalan (`Ca`), Galician (`Gl`), and Portuguese (`Pt`). Unlike these languages, Google has already launched voice search for Spanish which allows us to utilize a large amount of collected data. Therefore, we make use of unsupervised training techniques to enhance the current Spanish system. In addition, we provide a close examination of the `Es` embedded ASR system, which runs locally on the device. This `Es` embedded system is, to our knowledge, the first commercial large vocabulary ASR service that runs locally within the device, although it has obvious restrictions of CPU, memory and storage. Yet, it supports, without Internet connection, services such as Voice IME (which complements the device's tactile keyboard with an ASR system) and Dictation (which provides instantaneous, real-time ASR results).

We describe the challenges of serving queries that have a large vocabulary size. Our vocabulary contains often more than one million tokens since it covers URLs, names, locations, business names, etc. We also consider the challenges related to the data acquisition and annotation process for both supervised and unsupervised data, along with the issues related to the peninsula's multilingualism (e.g. Spanish queries as input to the `Eu`, `Ca` or `Gl` systems, or English queries as input to any of the 5 systems).

The paper is organized as follows. First, we describe our data collection in Section 2. Our system is described in Section 3, including acoustic, pronunciation and language models. Section 4 overviews the shrinking criteria for the `Es` embedded system. In Section 5 we show our experimental results on the five lan-

guages, our cross-languages comparisons and the performance of the embedded `Es` system. We conclude in Section 6.

2 Data Acquisition

2.1 Supervised Collection `Ca`, `Eu`, `Gl` and `Pt`

In developing large-scale commercial ASR systems, it is crucial that an efficient collection of high-quality data is identified. However, publicly available corpora are often scarce, due to legal limitations and the nature of the data itself. For our data acquisition, we utilized the following guidelines: phonetic differences should be captured by recordings from different geographical locations; for an efficient cross-lingual comparison, data transcriptions should be verified and classified based on language; to avoid channel mismatch, several types of mobile phones should be used in the recording of the audio; our data should be recorded matching real conditions (e.g., rooms, streets, shopping centers, cars, etc.); and the semantic nature of the utterance should be specific to the task (e.g. we prefer audio from voice queries).

For all of the aforementioned reasons, we decided to collect our own corpora. We collected spoken utterances in different acoustic environments from a variety of speakers [3]. Our data collection application displays prompts on a mobile device based on common and localized search queries. We recruited native speakers of each of the five Iberian languages and recorded more than 150,000 spoken queries per language. Our speakers included both female and male volunteers from different age ranges. We recorded the audio samples in both quiet and noisy environments, including in offices, in shopping centers, on public transportation and others. We used most of the data for training purposes and reserved approximately 15,000 utterances from each language for testing. The sets of speakers appearing in the training and testing sets were disjoint. Further details are included in Table 1.

2.2 Unsupervised Collection in `Es`

Despite the strong restrictions that were imposed on the collection of supervised data, human transcription accuracy represents a key limitation to improving the system's performance[4]. On the one hand, transcription is a task that requires expertise as well as specific knowledge about context and semantic information. This is true even if the transcription task is presented as a verification task (i.e. acceptance vs rejection of utterances, Section 2.1 above). On the other hand, it is often impractical to manually transcribe the amount of data used to train commercial systems. Even ideally transcribed collections often suffer from limitations (e.g. the number of speakers, devices or ages), which are hard to control in practical circumstances.

Therefore, for the `Es` system, we adopted an end-to-end closed-loop process that requires no external human transcription. Instead, it relies on real traffic

Lang.	Set	# Spkr.	# Queries	Duration	Type
Ca	train	687	204K	392 hrs	MC, M
	test	29	14.5K	27 hrs	MC, M
Eu	train	388	84K	188 hrs	MC, M
	test	43	14.7K	31 hrs	MC, M
Gl	train	429	128K	262 hrs	MC, M
	test	29	14.5K	28 hrs	MC, M
Es	train	>2000K	∼2000K	>2000 hrs	RT, A
	test	25K	∼25K	53 hrs	RT, M
Pt	train	763	153K	331 hrs	MC, M
	test	24	12K	26 hrs	MC, M

Table 1. Train and Test Data for All Iberian Languages Data can be of type manually collected data (MC), real traffic data (RT), manually transcribed data (M) or automatic transcribed data (A)

samples provided by our initial Es Voice Search service and a confidence model. The confidence model estimates the confidence posteriors of each utterance transcribed and filters them according to a threshold. This includes a $\sim$ x3000 factor in the number of speakers, a $\sim$ x10 factor in the hours of speech and a $>$ x3 factor in the number of queries. Further, a flat start training of the system was used. Our testing corpora were random samples of real traffic that had been manually transcribed. As a consequence these corpora are more challenging than those derived from data collections, but at the same time more representative of the true Word Error Rate (WER) expected in production. See Table 1 for details of our training and testing data sets for all languages. In all experiments (Section 5), the confidence model used is a logistic regression classifier trained over the word lattice posteriors. Its threshold was set to select approximately the top 20%utterances.

3 ASR System Description

3.1 Acoustic Modeling

The front-end is a 13-dimensional PLP coefficient with cepstral mean normalization and energy-based endpointer, which removes excessive silence. Each frame is spliced together with four preceding and four succeeding frames and then Linear Discriminant Analysis (LDA) is performed to yield 39-dimensional feature vectors.

The Acoustic Models (AMs) in our systems utilize a standard Decision Tree (DT) state clustering approach, which uses 3-HMM-state context-dependent (triphone) models and a variable number of Gaussians per state. Our AMs are gender-independent, maximum-likelihood trained interleaved with estimation of a global semi-tied covariance (STC) transform [5], followed by boosted MMI (BMMI) training [6].

For decoding, a statically constructed context-dependent lexicon and language model are used (see below), which are expressed as finite-state transducers and composed on-the-fly [7]. [1]

3.2 Language Modeling

In Voice Search, users speak their search queries and the system returns search results. It has been observed that speakers formulate their spoken queries similarly to the way they type them. This observation is consistent across languages. Therefore, the best available source of training data for our Language Models (LM) is deemed to be the records of the anonymized typed search queries. Since textual data varies across languages, we trained separate LMs on anonymized search query logs from the corresponding local search engine. For example, the `Eu` LM was trained on the textual queries directed to *http://www.google.es/#hl=eu*.

As in various regions of the Iberian peninsula, the Spanish language coexists with a native language. We find a high percentage of Spanish queries in the logs directed to the Basque, Catalan and Galician search engines. These queries mostly contained locations, names, brands, business names and URLs. Hence, we concluded that a voice search system that only supports one of these native language would be impractical for users; conversely, a system in Spain that only handles Spanish would be unsatisfactory for large portions of the population. Therefore, it was deemed very important that our `Ca`, `Eu` and `Gl` systems have the capability of supporting both native and Spanish queries. Similarly, since a significant number of English queries are directed to the local search engines of these Iberian languages, it is useful that our systems support common English-language queries as well.

For the aforementioned reasons, data from queries containing both frequent Spanish words in the `Ca`, `Eu` and `Gl` corpora, as well as English words in all language corpora are used for training the LMs. The LM for each language is trained using search queries sampled only from one year's worth of search query logs. In this work, we make use of 5-gram back-off LMs, trained with Katz smoothing [9] and entropy pruning [10]. Further details of the LM that we used for Voice Search can be found in [11]

3.3 Lexicon

For languages with complex letter-to-sound mappings, pronunciation dictionaries including all common words are typically written by experts. Fortunately, for all Iberian languages, the relationship between orthography and pronunciation is relatively regular and well understood. Hence, when a word is not found in the dictionary, its pronunciation(s) can be inferred by a number of rules to generate phoneme sequence(s).

However, to be able to map native words to their true phonemic representations, the words must first be classified as belonging to one of the following

[1] This system is similar to our Cantonese voice search system [8].

categories: *Native*, *Number*, *Composed* (e.g. URLs, emails) or *Non-Native*. Each category is then processed independently using the following rules.

- *Native* words in the system vocabulary use the pronunciation found in the dictionary, if available. Otherwise, the pronunciation is generated using native phonetic rules. When these fail, letter name pronunciations are generated for each character, which allow us to deal with acronyms.
- *Numbers* are first converted to a list of words using a number grammar (implemented with a finite-state transducer, FST), and those words are then processed as *Native*.
- *Non-Native* words in the LM are assumed to be English and are imported from the American English lexicon. Then, we map the English phonemes to the closest native phonemes, which allows us to model locally accented English speech without increasing the size of the phoneme inventory.
- *Composed* words are first converted as a sequence of words using a small dictionary of delimiter characters. Each token is then classified and processed based on its category.

Lexicons in `Ca`, `Eu` and `Gl`: `Ca` and `Eu` lexicons follow a variant of the method described above, which we denote as *dual lexicon*. In this case, the *Non-Native* category is processed using the `Es` lexicon and the corresponding Spanish-to-native phonetic mapping. This architecture covers all possibilities of native, Spanish and English words. Note that English words are processed as *Non-Native* words in the Spanish lexicon. This architecture is scalable to as many languages as are needed.

Without significant performance loss, additional simplification is made in the `Ca`, `Eu` and `Gl` lexicons. Specifically, the `Gl` lexicon used in our system is the same as the `Es` lexicon. The `Eu` lexicon uses native number grammars and phonetic rules inherited from the `Es` lexicon, but removes diacritics. Finally, the `Ca` lexicon uses specific number grammars and a pronunciation dictionary with 4 million entries that covers all common native words. It is also backed up with Spanish phonetic rules.

4 The `Es` embedded system

Our `Es` embedded system is conceptually devised in a way that is identical to that of the standard `Es` system. Yet, the challenge is to reduce the size of the final model and its number of parameters, while increasing the real time factor in a way that does not significantly degrade performance. Its final size is determined by the memory, CPU and space requirements of most portable devices running Android.

Both the AMs and LM are reduced in size to produce the embedded recognizer. The LM is shrunk to $\sim 3\%$ of its original size by using entropy pruning [11]. We also reduce the vocabulary size of the speech recognizer from around 1

Sys.	WER %	Sub %	Del %	Ins %	RT50	RT90
Ca	21.6	2.0	4.5	15.1	0.50	0.96
Eu	18.7	3.2	2.8	12.7	0.36	0.77
Gl	11.5	1.3	2.1	8.2	0.27	0.52
Pt	19.3	3.2	3.3	12.7	0.51	0.99
Es	30.6	5.7	3.5	21.3	0.44	0.97

Table 2. Overall performance in terms of WER. Real Time (RT) factors are measured excluding the 50% (RT50) and the 10% (RT90) of the slowest queries to make this measurement more robust to non-speech (garbage) inputs.

million words to about 50 thousand. The AM is reduced in size from 5K context dependent (CD) states and 214K Gaussians to 1K CD states and 127K Gaussians. This smaller AM size is accomplished by manipulating state occupancy thresholds during the context-dependency decision tree learning process, as well as by adjusting Gaussian learning parameters when performing distribution splitting during Expectation Maximization training of Gaussian mixture acoustic models. With these size adjustments to the model, we achieve a $\sim$ x3 speed up factor.

The resulting overall size of the Es model package containing the FST-based decoder graph combining the context dependency transducer, and the pronunciation and language models, as well as the acoustic model parameters, is 22MB. Because it is not feasible to distribute recognizers for all supported languages with the Android OS, this model package is offered as an additional download. This option is available to Android users as of the 4.1 "Jelly Bean" release.

5 Experiments

5.1 Voice Search Performance in Ca, Eu, Gl, Pt and Es

Table 2 shows the overall WERs of our five recognizers. Analyzing this table, (1) we achieve acceptable WERs for all evaluated languages, with similar WERs to those reported for other languages in Voice Search systems [12], [8], [2]. (2) We obtain a low WER for Gl; however, we believe that it is due, in part, to the particularly clean test set. It is important to note that we cannot directly compare the WER of Es system to the WERs of the other four systems because the Es test set has been collected using real traffic samples, which represents a more challenging task. All systems run below real time (RT90), which is a design criterion of our Voice Search systems.

5.2 Cross-Language Experiments

Cross-language experiments were conducted to evaluate the degree of support that could be achieved for the bilingual regions of Spain. In particular, given

System\Test Set	Ca	Eu	Gl	Pt	Es
Ca	21.6	99.5	65.4	91.4	**45.2**
Eu	85.3	18.7	62.8	101.1	**51.7**
Gl	75.2	101.3	11.5	95.3	**40.8**
Pt	102.2	121.5	93.6	19.3	90.4
Es	75.8	96.0	44.2	108.3	30.6

Table 3. WER (%) of the Cross-Language Evaluation

the number of Spanish queries found in our Ca, Eu and Gl data sets, we pay particular attention to how well the Spanish language is supported by these three systems.

Table 3 shows that the Ca, Eu and Gl systems evaluated over the Spanish test set achieve a 45.2%, 51,7% and 40.8% WER respectively. Interestingly, these WERs seem to be low given the difficulty of the Spanish test set and particularly when comparing them to their symmetric values in Table 3. Note that, for example, the performance of the Es system degrades to 96% WER when it is evaluated on the Basque test set.

Figure 2 reflects a differential analysis of the five presented systems. It is a graphical analysis of the values included in Table 2 that results from a Multi-Dimensional Scaling (MDS) [12]. As it is qualitatively to be expected, it shows that Eu and Pt are very distant, while Es and Gl are much closer systems. This finding suggests a potential combination the Es and Gl systems.

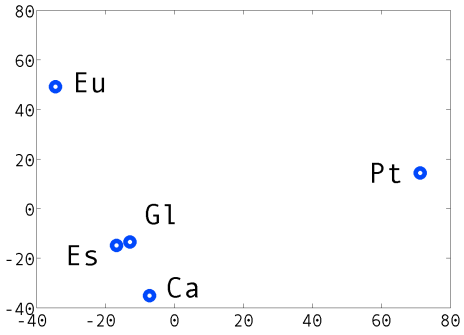


Fig. 2. Data-Driven Language Distance Map Derived by MDS from Cross-Language WER

System	WER (%)	RT90	% Size	Gauss	States
Uns. `Es`	30.6	0.97	100	214K	5K
Seed `Es`	34.5	0.91	$\sim$ 100	$\sim$ 214K	$\sim$ 5K
Emb. `Es`	40.3	0.37	3.2	127K	1K

Table 4. Unsupervised, Seed and Embedded `Es` systems performance

5.3 Seed, Unsupervised and Embedded `Es` Systems

We mentioned in Section 1, that Google has already launched Voice Search for Spanish. Here we will refer to this system as *seed* system. The *seed* system was used to acquire training data for our new *unsupervised* system (Section 2.2) and later, the model size of the *unsupervised* system was reduced to generate our new *embedded* system (Section 4). In this section we compare the performance of these three versions of the `Es` system, as it can be useful to illustrate the performance achieved by our shrinking, and unsupervised training strategies (Table 4).

On the one hand, the improvement derived mostly by our unsupervised training strategy is above 10% in WER and highlights the relevance of using data from real conditions, even when automatic transcriptions may have a large number of errors compared to the manual ones. On the other hand, the WER of the *embedded* system, which is 3% of the size of the *unsupervised* system and $\sim x3$ times faster, degrades $\sim$ 33% in WER with respect to the *unsupervised* system. We believe that this degradation is relatively small given the complexity of the task and the *embedded* system can still achieve a good user experience.

6 Conclusions

This paper introduces Google's Voice Search systems for the official languages spoken in the Iberian peninsula: Basque, Catalan, Galician, Portuguese and Spanish. These systems have the potential to reach more than 55 million people. We reviewed the elements that integrate their language models, acoustic models and lexicons. In particular, we described how the existing multilingualism in Spain can be addressed using combined lexicon strategies. For each language, figures of performance were provided in terms of WER and Real Time Factor. We also conducted cross-language experiments, during which we trained on one language and tested on the others. When projecting these cross-language WERs into 2D space using multidimensional scaling, the resulting map illustrates the degree of similarity across systems. This work also analyzed the challenges associated with the unsupervised training used in the `Es` system. Finally, we described the implementation details and performance of the new embedded `Es` system.

References

1. M. Bacchiani, F. Beaufays, J. Schalkwyk, M. Schuster, and B. Strope, "Deploying goog-411: Early lessons in data, measurement, and testing," in *In Proceedings of ICASSP*, 2008, pp. 5260–5263.
2. J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Garrett, and B. Strope, "Google Search by Voice: A Case Study," in *Advances in Speech Recognition: Mobile Environments, Call Centers, and Clinics*, A. Neustein, Ed. Springer-Verlag, 2010.
3. T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Interspeech*, 2010.
4. P. W. L. Wang, M. Gales, "Unsupervised training for mandarin broadcast news conversation transcription," in *In Proceedings of ICASSP*, vol. 4, 2007, pp. 353–356.
5. M. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.
6. D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature space discriminative training," in *ICASSP*, 2008.
7. C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *9th Intl. Conference on Implementation and Application of Automata*, 2007, `www.openfst.org`.
8. Y.-H. Sung, M. Jansche, and P. Moreno, "Deploying Google Search by Voice in Cantonese," in *Interspeech*, 2011.
9. S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. ASSP*, 1987.
10. A. Stolcke, "Entropy-based pruning of backoff language models," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
11. C. Chelba, J. Schalkwyk, T. Brants, V. Ha, B. Harb, W. Neveitt, C. Parada, and P. Xu, "Query language modeling for voice search," in *Proceedings of the 2010 IEEE Workshop on Spoken Language Technology*, 2010, pp. 127–132.
12. F. Biadsy, P. Moreno, and M. Jansche, "Google's cross-dialect arabic voice search," in *In Proceedings of ICASSP*, 2012, pp. 4441–4444.

transLectures

J. A. Silvestre-Cerdà, M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan

Universitat Politècnica de València (UPV), Camí de Vera s/n, 46022 València
{jsilvestre,mdelagua,ggarces,ggasco,agimenez,aperez,isanchez,nserrano rnadine,jvalor,jandres,jcivera,josanna,ajuan}@dsic.upv.es
http://translectures.eu

Abstract. transLectures (Transcription and Translation of Video Lectures) is an EU STREP project in which advanced automatic speech recognition and machine translation techniques are being tested on large video lecture repositories. The project began in November 2011 and will run for three years. This paper will outline the project's main motivation and objectives, and give a brief description of the two main repositories being considered: VideoLectures.NET and poliMedia. The first results obtained by the UPV group for the poliMedia repository will also be provided.

Keywords: language technologies, machine translation, automatic speech recognition, massive adaptation, intelligent interaction, education, video lectures, multilingualism, accessibility, opencast matterhorn

1 Introduction

transLectures is the acronym adopted for the EU (FP7-ICT-2011-7) STREP project entitled "Transcription and Translation of Video Lectures". It began on 1 November 2011 and will run until 31 October 2014. The transLectures consortium includes video lecture providers (users), experts in automatic speech recognition (ASR) and machine translation (MT), and professional transcription and translation providers:

UPV Universitat Politècnica de València, València, Spain.
XEROX Xerox S.A.S., Grenoble, France.
JSI Jozef Stefan Institute, Ljubljana, Slovenia.
RWTH Rheinisch-Westfaelische Technische Hochschule, Aachen, Germany.
EML European Media Laboratory GmbH, Heidelberg, Germany.
DDS Deluxe Digital Studios Ltd, London, UK.

The aim of this paper is to outline the motivation behind the transLectures project (sec. 2) and the main objectives established for it. Brief descriptions will be provided of the poliMedia and VideoLectures.NET repositories (sec. 3), as well as of the first results obtained by UPV for poliMedia (sec. 4).

2 Motivation and project objectives

Online multimedia repositories are growing rapidly and becoming evermore consolidated as key knowledge assets. This is particularly true in the area of education, where large repositories of video lectures are being established on the back of increasingly available and standardized infrastructures. A well-known example of this is VideoLectures.NET, a free and open access web portal that has already published more than 10,000 educational videos. VideoLectures.NET is a major player in the diffusion of the open-source Matterhorn platform currently being adopted by many education institutions and organizations within what is known as the Opencast community[1].

As with many other repositories, most of the video lectures available on VideoLectures.NET are neither transcribed nor translated, despite the clear need to make their content accessible to speakers of different languages and people with disabilities. Transcription and translation would also enable the incorporation of search and analysis functions, such as lecture classification, summarisation or plagiarism detection. However, a cost-effective solution that can do so to a reasonable degree of accuracy has yet to be developed.

The aim of the transLectures project is to do just this: to develop innovative, cost-effective solutions for producing accurate transcriptions and translations of the lectures published on VideoLectures.NET, with their compatibility and usability across other Matterhorn-related repositories informing all design decisions. Our starting hypothesis is that the gap that current automatic speech recognition (ASR) and machine translation (MT) technologies must bridge in order to achieve acceptable results for the kind of audiovisual collections being considered in the project is relatively small, and can be closed by pursuing the following three scientific and technological objectives:

1. *Improvement of transcription and translation quality by massive adaptation.* ASR has yet to reveal its full potential in the generation of acceptable transcriptions for large collections of audiovisual material. However, that potential is within reach and relatively little further research into ASR technology is required; rather we must learn to better exploit the wealth of knowledge we have at hand. More precisely, through this project we hope to demonstrate that transcriptions of a reasonable quality can be obtained through the massive adaptation of general-purpose models on the basis of lecture-specific variables, such as speaker, topic and, more importantly, time-aligned slides. Only once we have achieved acceptable transcriptions can we hope to address the adaptation of translation models with any degree of success.
2. *Improvement of transcription and translation quality by intelligent interaction.*
   Massive adaptation can and will deliver substantial contributions to the improvement of overall quality, but it is our belief that sufficiently accurate translations are unlikely to be achieved through fully-automated approaches alone: in order to reach the desired levels of accuracy, we must consider user

[1] See http://www.opencastproject.org/project/matterhorn

interaction. Current user models for the transcription and translation of audiovisual material tend to be batch-oriented. Under this model, an initial transcription/translation is generated by the system offline and then sent to the user to be post-edited manually without system participation. In our view, these models only give satisfactory results when highly collaborative users are working on near-perfect system output. Otherwise, a more intelligent interaction model is required that saves on user supervision and allows the system to learn from user supervision actions dynamically. In transLectures, our aim is to develop innovative, truly-interactive models in which the system learns from and reacts to each user supervision action as and when received. We are also exploring various modes of interaction, to reflect the multiple roles users can have.

3. *Integration into Matterhorn to enable real-life evaluation.*
   In contrast to many past research efforts in which system prototypes are evaluated in the lab alone and are largely inapplicable to real-life settings, we will be developing tools and models for use with Matterhorn. We will therefore be able to evaluate their usefulness using real-life data in a real-life context.

We are confident that the innovative solutions we develop in these three areas will be deployed rapidly across many educational repositories in Europe and worldwide, allowing these portals to overcome language barriers and reach wider audiences while supporting linguistic diversity. In transLectures, we will be testing our ideas on VideoLectures.NET and on a smaller repository of Spanish video lectures, poliMedia, which is also part of the Matterhorn community. For transcription, we are considering English and Slovenian through VideoLectures.NET and Spanish through poliMedia, with the former accounting for more than 90% of all lectures. Meanwhile, for translation we are considering the language pairs {Spanish, Slovenian} into English, and English into {French, German, Slovenian, Spanish}.

3 VideoLectures.NET and poliMedia

The best proof of the effectiveness of the transLectures approach would be the cost-effective production of accurate transcriptions and translations within the context of one or more video lecture repositories. As discussed above, we are testing our ideas on two case studies: VideoLectures.NET and poliMedia.

VideoLectures.NET was founded in 2001 as an internally-funded project and is now run by the dedicated Centre for Knowledge Transfer in Information Technologies at the Josef Stefan Institute (JSI), Ljubljana, Slovenia. It is a free and open access repository of video lectures mostly filmed by people from JSI[2] at major conferences, summer schools, workshops and promotional events from many scientific fields. Indeed, VideoLectures.NET is being used as an educational platform for several EU-funded research projects[3]; and by various open education

[2] See promo video at http://videolectures.net/april2010_videolectures_promo
[3] See http://videolectures.net/eu_supported

resource organizations such as the OpenCourseWare Consortium, MIT OpenCourseWare and Open Yale Courses; as well as other scientific institutions like CERN. In this way, VideoLectures.NET brings together high-quality educational content that has been recorded to similarly high quality standards. All lectures, accompanying documents, information and links are systematically selected and classified through the editorial process, and user comments are also taken into account. Video editing is carried out in-house and is completely uncensored, in the sense that lectures are never edited in a way that would allow content or viewer manipulation. Most lectures are also accompanied by time-aligned presentation slides[4].

poliMedia is a recent, innovative service for the creation and distribution of multimedia educational content at the UPV. It is designed primarily to allow UPV professors to record their courses in video blocks lasting up to 10 minutes. As in VideoLectures.NET, the videos are accompanied by time-aligned slides. However, in contrast to VideoLectures.NET, the video recordings are filmed at specialized studios under controlled conditions to ensure maximum recording quality and homogeneity. Professors are filmed against a constant-colour background in order to be able to post-produce presentations in which the professor appears, properly scaled, against a backdrop of the corresponding slides. Please visit http://polimedia.blogs.upv.es/?lang=en for more details and examples[5].

transLectures is focusing on English and Slovenian video lectures from VideoLectures.NET and on Spanish video lectures taken from poliMedia. For an idea as to the complexity of the task, some basic statistics have been provided in Table 1.

	English	Slovenian	Spanish	Total
Authors	6900	1347	734	8981
Lectures	9720	1103	5056	15879
Avg. lecture duration (min)	45	45	9	34
Transcribed lectures	111	0	7	118
Lectures with slides	7013		734	

Table 1. Basic statistics on the video lectures considered in transLectures.

4 First UPV results on poliMedia

The UPV obtained its first results for poliMedia working from a set of 115 hours of video lectures manually transcribed using the Transcriber tool [1]. From this set, a standard partition was defined with three speaker-independent sets: training, development and test. This will allow ongoing scientific evaluation through-

[4] e.g. http://videolectures.net/translingeu2010_uszkoreit_rossi_welcome

[5] See http://polimedia.upv.es/visor/?id=39f62a9a-4cf5-bd4e-92f3-cb34e4792a85 for a brief presentation in Spanish, with subtitles available in English.

out the project. Table 2 shows the basic statistics for this standard partition.

	Training	Development	Test
Videos	559	26	23
Speakers	71	5	5
Hours	99h	3.8h	3.4h
Sentences	37K	1.3K	1.1K
Vocabulary	28K	4.7K	4.3K
Running words	931K	35K	31K
OOV words	-	4.6%	5.6%
Perplexity	-	222	235

Table 2. Statistics for the standard partition for scientific evaluations on poliMedia.

The baseline UPV ASR system is based on the RWTH ASR system [2, 4] for acoustic modelling and the SRILM toolkit [6] for language modelling. The RWTH ASR system includes state-of-the-art speech recognition technology for acoustic model training and recognition. It also includes speaker adaptation, speaker adaptive training, feature extraction for audio files, unsupervised training, a finite state automata library and an efficient tree search recognizer. For its part, the SRILM toolkit is a well-known language modelling toolkit used across different natural language applications.

The audio data was extracted and preprocessed from the videos, in order to then extract the mel-frequency cepstral coefficients (MFCCs) [5]. Then, monophoneme and triphoneme acoustic models were trained by adjusting different parameters such as the number of states, Gaussian components, leaves of the triphoneme CART, etc. in the development set. The lexicon model was obtained by applying phonetic transliteration to all of the training vocabulary words. An n-gram language model was trained on the transcribed text after filtering out functional symbols such as punctuation marks, silence annotations, etc. Meanwhile, external resources were used to enrich the in-domain language model. Specifically, we considered the linear combination of our in-domain language model with a large, external out-of-domain language model computed from the Google n-gram corpus [3]. A single parameter λ governs the linear combination of the poliMedia language model and the Google n-gram model, which is optimized in terms of perplexity on the development set.

The entire system, including the acoustic, lexicon and language models, was trained on the poliMedia training set. The ASR system parameters were optimized in terms of word error rate (WER) on the development set. A significant improvement of more than 5 WER points was observed when moving from monophoneme to triphoneme acoustic models. Triphoneme models were inferred using the conventional CART model using 800 leaves. In addition, other parameters obtained to train the best acoustic model included 2^9 components per Gaussian mixture, 4 iterations per mixture, and 5 states per phoneme. The in-domain

language model was an interpolated trigram model with Kneser-Ney discount. Higher and lower order n-gram models were also assessed, but no better performance was observed.

From the triphoneme ASR system, several refinements to the language model were evaluated. The in-domain language model trained on the poliMedia corpus was interpolated with the out-of-domain Google n-gram corpus [3]. These two language models were interpolated in order to minimize perplexity in the development set, using an approximate λ value of 0.65 for the in-domain language model and of 0.35 for the out-of-domain language model. Two interpolations were performed using different vocabulary sets, the first containing only vocabulary matching poliMedia ("LM Interpolation") and a second made up of the poliMedia vocabulary plus the 50,000 most frequent words in the Google n-gram corpus ("Google 50K"). The final experimental results in terms of WER in the test set are shown in Table 3.

System	WER	OOV
Monophoneme Model	44.6	5.6%
Triphoneme Model	39.4	5.6%
+LM Interpolation	34.6	5.6%
+Google 50K Vocab	33.7	3.5%

Table 3. Test-set WER for several ASR system refinements.

As shown in Table 3, there is a significant improvement by 5.7 WER points over the triphoneme system when the language model interpolated with the "Google 50K Vocab" vocabulary set is applied. As expected, the decrease in WER is directly correlated with the number of out-of-vocabulary words (OOVs) in the test set, since the Google n-gram corpus provides a better vocabulary coverage. A similar trend is observed when comparing perplexity figures for the triphoneme system with those observed for the "LM Interpolation" system. Specifically, perplexity drops significantly from 235 to 176 simply by interpolating our in-domain language model with the Google n-gram language model containing poliMedia vocabulary alone.

5 Concluding remarks

In this paper, we have outlined the transLectures project's main motivation and objectives, and given a brief description of the two video lecture repositories being considered: VideoLectures.NET and poliMedia. We have also reported the first ASR results obtained by UPV for poliMedia.

Our current work is aimed at the application of massive adaptation techniques on the basis of lecture-specific knowledge and, in particular, time-aligned slides. We are also running similar experiments on the larger VideoLectures.NET repository.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755. Funding was also provided by the Spanish Government (iTrans2 project, TIN2009-14511; FPI scholarship BES-2010-033005; FPU scholarship AP2010-4349).

References

1. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication special issue on Speech Annotation and Corpus Tools 33(1–2) (2000)
2. Lf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlter, R., Ney, H.: The rwth 2007 tc-star evaluation system for european english and spanish. In: Proc. of Interspeech. pp. 2145–2148 (2007)
3. Michel, J.B., et al.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182.
4. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lf, J., Schlter, R., Ney, H.: The rwth aachen university open source speech recognition system. In: Proc. of Interspeech. pp. 2111–2114 (2009)
5. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition 54(4), 543–565 (2012)
6. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proc. of ICSLP (2002)

META-NET. Official Languages of Spain in the Digital Age

Asunción Moreno[1], Núria Bel[2], Maite Melero[3], Carmen García-Mateo[4], Inma Hernáez[5], Sergio Oller[6], Aljoscha Burchardt[7], Kathrin Eichler[7], Georg Rehm[7], and Hans Uszkoreit[7]

[1] Universitat Politècnica de Catalunya. Centre TALP
asuncion.moreno@upc.edu,
[2] Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
nuria.bel@upf.edu,
[3] Barcelona Media
maite.melero@barcelonamedia.org
[4] Universidade de Vigo. AtlantTIC Research Center
carmen.garcia@uvigo.es
[5] University of the Basque Country. Aholab Signal Processing Laboratory
inma.hernaez@ehu.es
[6] Universitat de Barcelona
sergioller@gmail.com
[7] DFKI GmbH, Berlin, Germany
{aljoscha.burchardt,kathrin.eichler,georg.rehm,hans.uszkoreit}@dfki.de

Abstract. In the framework of the Network of Excellence META-NET, a White Paper series has been created to chart the status of Language Technologies for 30 European languages. The exhaustive analysis reports on the state of a language including social and technological aspects and statistics about the availability of Language Resources and Tools. The White Papers are addressed to politicians, journalist and decision makers. In this paper, we show a comparative study of the findings on Language Technologies and Tools for the four official languages spoken in Spain.

Keywords: Language technologies, Basque, Catalan, Galician, Spanish

1 Introduction

META-NET [10] is a Network of Excellence funded by the European Commission through a cluster of coordinated projects: T4ME [12], CESAR [13], META-NORD [14], and METANET4U [15]. The network also has multiple unfunded members. The network currently consists of 60 members from 34 European countries [2]. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology, ranging from household electronics, machinery and vehicles to computers and robots.

Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action: META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a common Strategic Research Agenda (SRA) [1]. Its main focus is to build a coherent and cohesive LT community in Europe, by bringing together representatives from extremely diverse and highly fragmented groups of stakeholders. Of great importance is the White Paper series [4] produced for more than 30 languages, of which the main results for the official languages spoken in Spain are compiled in this paper.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and Web services that are documented with high-quality metadata and organised in standardised categories [3]. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for the members of the community.

As of November 2012, META-NET consists of 60 research centres from 34 European countries. META-NET is working with stakeholders from economy (software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

In the framework of META-VISION, a White Paper series has been created to chart the status of Language Technologies for 30 European languages.

The exhaustive analysis reports on the state of each language including social and technological aspects, and statistics about the availability of Language Resources and Tools. In this paper we show a comparative study of the findings on Language Technologies and Tools for the four official languages spoken in Spain. Next section briefly describes the contents of the White Papers. Section 3 shows a detailed comparison of the language technologies, applications and solutions for Basque, Catalan, Galician and Spanish. Section 4 show the results of this analysis among the other analyzed European languages. Conclusions are included in section 6.

2 White Paper series

META-VISION has conducted an analysis of current language resources and technologies in a White Paper series. The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe [11]. More than 200 experts from all over Europe have contributed to the 31 volumes of the White Paper series and more than 8,000 printed copies were disseminated to key decision makers, politicians and journalists in September and October of 2012. Four White Papers for the four official languages spoken in Spain: Basque [7], Catalan [5] Galician [8] and Spanish [6] were produced within a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others.

Each volume has three main parts; first of all, there is a description of how the diversity of languages in Europe is a rich cultural heritage and, at the same time, a risk for the survival of the languages. The White Paper series show that Language technologies can play a key role helping multilingual societies to survive. The second part of the White Papers deals with the specificities of each language and how it is promoted at national and international level. The third part provides a general description of the Language Technologies, with a presentation of the State of the Art, and also the programmes, projects and efforts at national and international levels done to improve the Language Technologies and Resources for each particular language. This part ends with a summary of availability of tools and resources and a cross-language comparison between the intended language and the other 29 European languages included in the White Paper series.

The analysis of the White Paper series shows that the availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research. In the specific case of Spain, we find one official language, Spanish, and three co-official languages: Catalan, Galician and Basque. Spanish

Fig. 1: Compilation of White Papers

is, in addition, an official language in Europe, most of the countries in Latin America and is broadly spoken in the US. Preserving multilingualism in Spain has not been an easy task. It is the result of a complex process to intentionally preserve cultural and linguistic identity within and among the various regions and peoples of Spain. Similar to the use of English in the European case, direct communication between citizens of different language areas of Spain, often need to use Spanish as a lingua franca.

3 State of Language Technology in the four official languages of Spain

A Table with a summary of the current state of language technology (LT) support for each of the languages was elaborated. For each intended language, the

data were taken from a survey among a representative number of researchers and industrial people in the LT community. Two main areas were considered: Language Technology including tools, technologies and applications; and Language Resources including resources, data and knowledge data bases. A number of representative Language Technologies were chosen for the study: Speech Recognition, Speech Synthesis, Grammatical analysis, Semantic analysis, Text generation, and Machine Translation. For the Language Resources study, the following groups were chosen: Text corpora, Speech corpora, Parallel corpora, Lexical resources and Grammars.

Experts were asked to rate in a scale from 0 (very low) to 6 (very high) the existing tools and resources for each group according to seven criteria: Quantity, Availability, Quality, Coverage, Maturity, Sustainability, and Adaptability.

Figures 2 and 3 show the mean scores obtained for each language in each Language Technology for the criteria Quality and Coverage. Figures 4 and 5 show the mean scores obtained for each language in each Language Resources group for the criteria Quality and Coverage

The key results of the White Paper series for the official languages of Spain can be summed up as follows:

- Speech processing currently seems to be slightly more mature than the processing of written text. In fact, for Spanish and Catalan, speech technology has already been successfully integrated into many everyday applications, from spoken dialogue systems and voice-based interfaces to mobile phones and car navigation systems.
- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics. The more semantics a tool takes into account, the more difficult is to find the right data; more efforts to support deep processing are needed.
- Research has successfully led to the design of medium to high quality software for basic text analysis, such as tools for morphological analysis and syntactic parsing. But advanced technologies that require deep linguistic processing and semantic knowledge are still in their infancy.
- As to resources, there are interesting resources in Spanish and Catalan, but they are not easily available for research. There are a number of corpora annotated with syntactic, semantic and discourse structure mark-up, but again, there are not nearly enough language corpora containing the right sort of content to meet the growing need for more deep linguistic and semantic information.
- In particular, there is a lack of the sort of parallel corpora that form the basis for statistical and hybrid approaches to machine translation. Parallel corpora exist, between Spanish and English, and between Spanish and other languages from Spain. However, parallel corpora between the four languages and other foreign languages are mostly missing.
- Many of these tools, resources and data formats do not meet industry standards and cannot be sustained effectively. A concerted programme is required to standardise data formats and APIs.

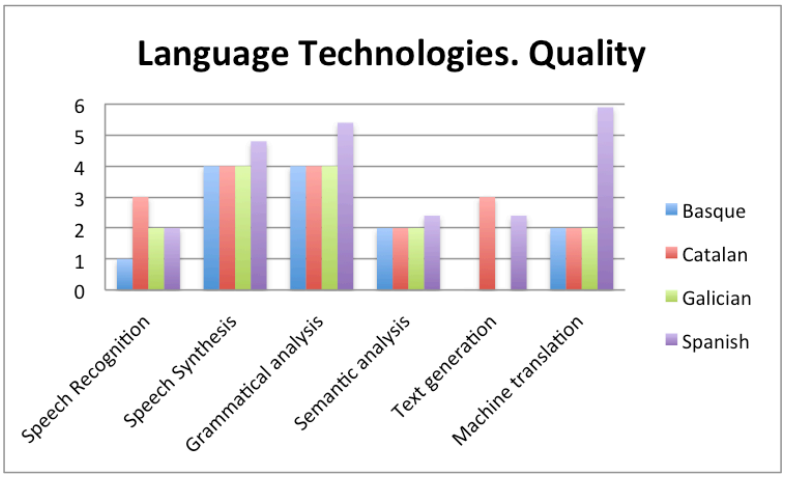


Fig. 2: Quality scores for Language Technologies

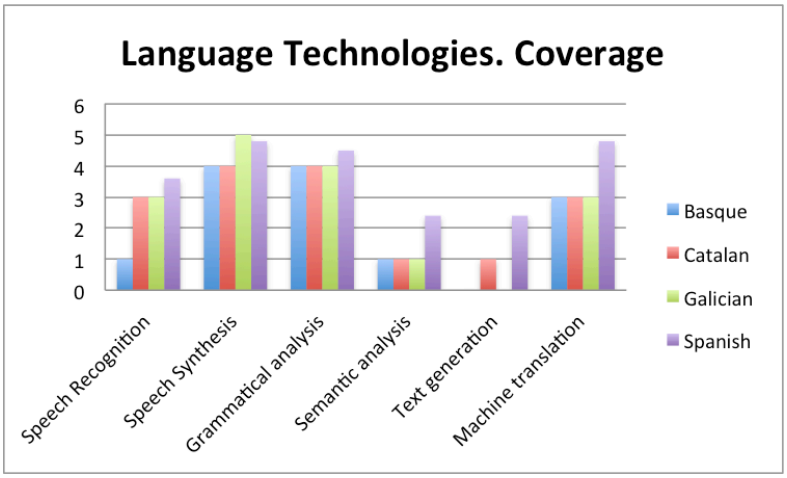


Fig. 3: Coverage scores for Language Technologies

– An unclear legal situation restricts making use of digital texts, such as those published online by newspapers, for empirical linguistic and language technology research, for example, to train statistical language models. Together with politicians and policy makers, researchers should try to establish laws or regulations that enable them to use publicly available texts for language-related R&D activities.
– The cooperation between the Language Technology community and those involved with the Semantic Web and the closely related Linked Open Data movement should be intensified with the goal of establishing a collaboratively maintained, machine-readable knowledge base that can be used both

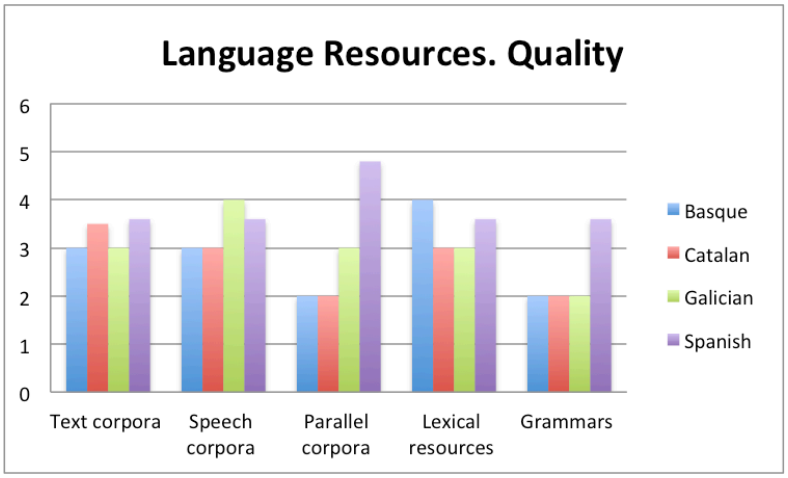


Fig. 4: Quality scores for Language Resources

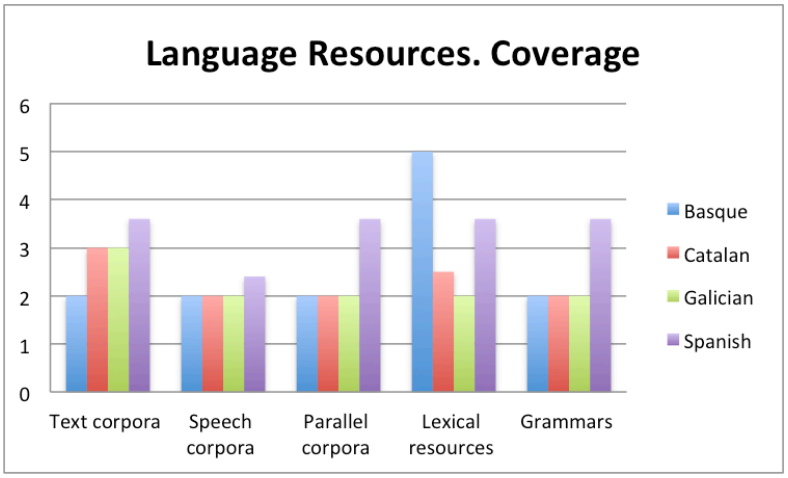


Fig. 5: Coverage scores for Language Resources

in web-based in-formation systems and as semantic knowledge bases in LT applications – ideally, this endeavour should be addressed in a multilingual way on the European scale.

To conclude, in a number of specific areas of language research, we have software with limited functionality available today. Obviously, further research efforts are required to meet the current deficit in processing texts on a deeper semantic level and to address the lack of resources such as parallel corpora for machine translation.

3.1 Cross-language comparison

The current state of LT support varies considerably from one language to another. In order to compare the situation between languages, this section presents an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using the following five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

The analysis shows that, thanks to large-scale LT funding in recent decades, Spanish is better equipped than most other languages. It compares well with most large languages, such as French and German. But LT resources and tools for Spanish clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications.

The results also show that, thanks to LT funding programmes from the Local and Central Governments in recent decades, the Basque, Catalan, and Galician languages are equipped as most of other European languages. They compare well with languages with a similar number of speakers such as Hungarian, Greek, and European Portuguese despite these are official languages of EU countries. But LT resources and tools for these languages clearly do not yet reach the quality and coverage of comparable resources and tools for the Spanish language, which is in a good position in almost all LT areas.

For speech processing, current technologies perform well enough to be successfully integrated into a number of industrial applications such as Interactive

Voice Response systems and constrained domain dictation systems. Machine Translation systems get a good performance, especially between the language pairs Spanish-English, Spanish-Catalan and English-Catalan. However, to build more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas, including high-quality machine translation.

4 Conclusions

The results of the study in the series of white papers show that there are serious differences in the support of European languages in the field of language technologies. While for some languages and certain applications, there is a good coverage in both software and language resources, for other languages, especially for minority languages, there are serious deficiencies in some areas.

In the case of Basque, Catalan, and Galician, the results of the study concerning the support of existing language technologies, are cautiously optimistic. There is a very active research community that has been working on research projects funded by the Local Governments and the Central Government. They have produced and distributed a number of language resources. Compared with other European languages, the resources and technology support is at a similar level to Greek, Hungarian or Norwegian being all official languages of European countries.

The range of products and applications available in Spanish greatly exceed those produced in the other three languages and the position of Spanish is comparable to French or German.

Language technologies are increasingly complex and require a large amount of data and resources to advance significantly. There are needs to coordinate the various sources of regional, national and European funding programs to circumvent language technologies support differences between the European languages.

5 Acknowledments

The authors of this document are grateful to the authors of the White Paper on German [9] for permission to re-use selected language-independent materials from their document. The authors acknowledge the support offered by the EU project T4ME [7], and the EU project Metanet4u [10]. Authors want to acknowledge their contribution to the coauthors and contributors of the White Paper series: for the Galician version, to the co-author Monsterrat Arza Rodríguez, and furthermore, to thank Dr. Xavier G. Guinovart (University of Vigo), Dr. Eduardo R. Banga (University of Vigo), Dr. Xosé Luis Regueira (University of Santiago de Compostela) and Mr. José Ramom Pichel (Imaxin Software) for their contributions to the Galician white paper. Material available at the web

pages of 'Consello da Cultura Galega Proxecto LOIA" and 'Secretaría Xeral de Política Lingüística Xunta de Galicia" was also been used. For the Catalan White Paper, we thank to the coauthors Eva Revilla, Emilia Garcia and Dr. Sisco Vallverdú, and to the contributors Dr. Lluís Padró, Dr. José B. Mariño, Dr. José R. Fonollosa and Dr. Climent Nadeu (all from Universitat Politècnica de Catalunya), Dra. Toni Martí (Universitat de Barcelona), Joan Soler (Institut d'Estudis Catalans), and Mercè Lorente (IULA-UPF). For the Spanish White Paper, authors thank his contribution to Dr. Toni Badia (UPF). For the Basque White Paper, authors thank his contribution to E. Navas (UPV/EHU), I. Odriozola (UPV/EHU) , K. Sarasola (UPV/EHU) , A. Daz de Ilarraza (UPV/EHU) , I. Leturia (Elhuyar Foundation), A. Daz de Lezana (Basque Government), B. Oihartzabal (CNRS), J. Salaberria (CNRS).

References

1. "Strategic Research Agenda for Multilingual Europe 2020", META Technology Council (eds.), 2012, META-NET, available at http://www.meta-net.eu/sra
2. Rehm G., and Uszkoreit H. Multilingual Europe: A challenge for language tech. MultiLingual, 22(3):5152, April/May 2011.
3. Federmann C., Giannopoulou I., Girardi C., Hamon O., Mavroeidis D., Minutoli S., Schrder M. META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey. 2012
4. META-NET White Paper Series: Europe's Languages in the Digital Age Georg Rehm and Hans Uszkoreit Eds., Spinger. 2012
5. Moreno A., Bel, N., Revilla E., Garcia, E., Vallverdú S. The Catalan Language in the Digital Age. Springer (2012)
6. Melero M., Badia T., Moreno A. The Spanish Language in the Digital Age. Springer (2012)
7. Hernáez I., Navas E., Odriozola I., Sarasola K., Daz de Ilarraza A., Leturia I., Daz de Lezana A., Oihartzabal B., Salaberria J. The Basque Language in the Digital Age. Springer (2012)
8. García-Mateo C., Arza, M. The Galician Language in the Digital Age. Springer (2012)
9. Burchardt A., Egg M., Eichler K., Krenn B., Kreutel J., Leßmöllmann A., Rehm G., Stede M., Uszkoreit H., Volk M. Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age. Springer 2012.
10. http://www.meta-net.eu
11. http://www.meta-net.eu/whitepapers
12. http://t4me.dfki.de/
13. http://www.meta-net.eu/projects/cesar
14. http://www.meta-nord.eu/
15. http://metanet4u.eu

Special Session
Demos

Hermes: a mobile distributed speech translation system

Emilio Granell, Carlos-D. Martínez-Hinarejos, and Vicent Tamarit

Instituto Tecnológico de Informática - Universitat Politècnica de València
Camino Vera s/n, 46022, Valencia, Spain
emgraro@posgrado.upv.es,cmartine@dsic.upv.es,vtamarit@dsic.upv.es

Abstract. In this work we present a distributed system for automatic interpretation. This system consists of a client application developed for mobile devices with Android operating system, which interacts with users, and a server dedicated to the automatic speech recognition and machine translation. The domain of interpretation is limited to a particular task based in the interaction that a tourist may have on arriving at a hotel. Our system is able to translate standard sentences in oral communication at the reception of a hotel from Spanish into English.

Keywords: Speech recognition, speech translation, distributed systems

1 Introduction

In today's world distances have shrunk and many borders are gone, as is the case on the European Union. Given these conditions, travelling is much easier and faster than in the last century and therefore every day language interpreters are more necessary. This fact is an opportunity for the scientific community regarding automatic speech recognition (ASR) and machine translation (MT) to develop applications and systems that allow communication between two people who do not speak a common language.

A remarkable work in automatic interpretation is the system of voice to voice machine translation for telephony EuTrans [1,2]. EuTrans is now entirely operative for two translation directions: Italian-English (*FUB* task) and Spanish-English (*tourist* task). The FUB task refers to oral communication in the telephone switchboard of a hotel, while the tourist task refers to that which may occur in the reception desk of a hotel.

Mobile devices are small devices with some processing capabilities, permanent or intermittent connection to a network, limited memory and designed specifically for a function, although they can perform other more general functions. With a mobile device that has processing capability, memory and network connection, we can use a client-server architecture to perform complex tasks that exceed the limits of its capabilities.

In this demonstration we present a system for automatic interpretation that makes communication possible between two people without a common language. Based on EuTrans, the domain of interpretation is limited to the *tourist* task (translate common phrases in oral communication in a hotel reception from Spanish into English). For the development of the architecture of this system we have relied on the ETSI 201 108 standard [3]. Thus, this system will consist of a server, responsible for conducting the recognition / translation, and a client application installed on a mobile device with which the user will interact.

Although there exist currently on the market interesting speech recognition and machine translation solutions for mobile devices, our system allows to train and adapt the acoustic and translation models for a specific task.

The rest of the paper is structured as follows. Section 2 offers the general design and the architecture of our system of automatic interpretation *Hermes*. Section 3 provides the conclusions and future work.

2 Hermes

Hermes is a distributed system with a client-server architecture. Hermes is composed of a client application that runs in a mobile device that communicates by sockets with a server, so that the mobile device only is responsible for the interaction with the user. This device digitises the user voice and shows to him/her the answer of the system, whereas the server performs the speech recognition and machine translation tasks, because they require more processing power and memory.

Figure 1 shows the operation of Hermes. As we can see, the client application in the mobile device saves the voice of the user in an audio file, sends this file to the server and shows the response received from the server to the user. On the other hand, the server keeps waiting for a request of the client application, and when it receives a new request, accepts and receives the audio file with the voice from the mobile device. Then, the server uses a speech recognition and translation engine on the received file, and its recognition result is returned to the mobile device through the socket connection.

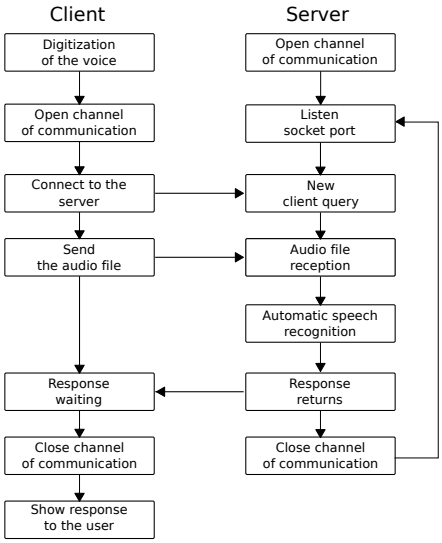
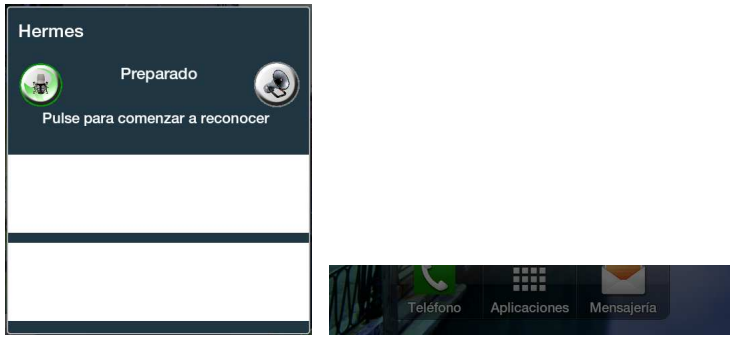


Figure 1. Model of operation of Hermes.



(a) Start screen. (b) Application menu.

Figure 2. Main activity.

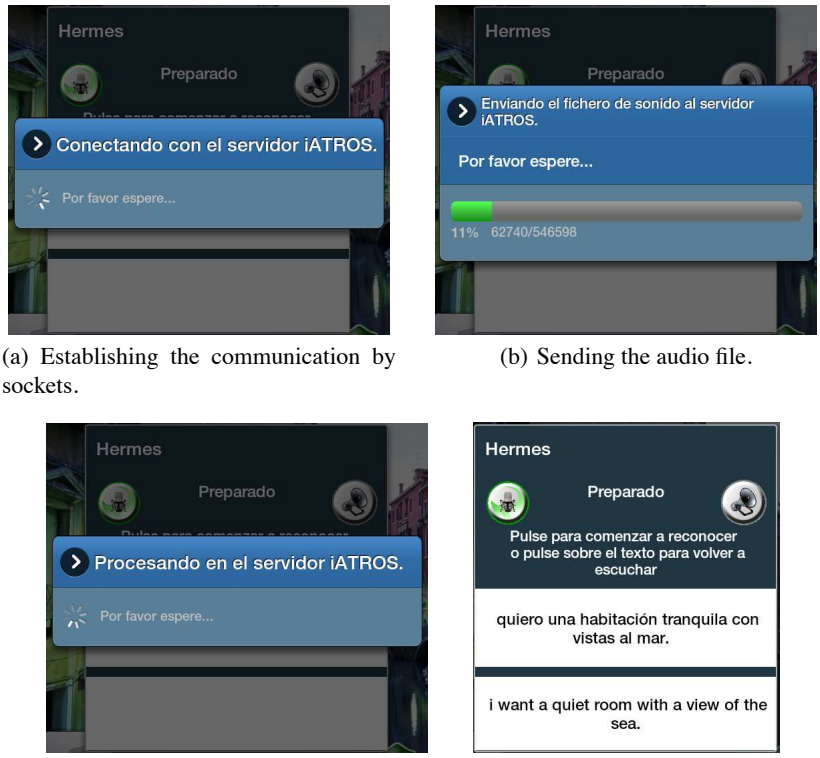
2.1 Client application

The client application was developed in Android [4,5,6] since it is a free platform and it has been very successful in the latter years. The mobile device elected for the development of this project was a tablet *Samsung Galaxy Tab*, that has a processor of 1 GHz, 16 GB of internal storage, 512 MB of RAM and Android 2.2. However, the application has been developed using the Android SDK with the API 7. This means that this application can be installed in mobile devices with Android from the version 2.1.

The client application is composed of three activities: main, digitisation of the voice, and communication with the server.

Activity 1: Main When the user initiates the application in his mobile device, he finds the start screen showed in the Figure 2(a). In this screen the user finds a green button with a microphone inside, a grey button with a speaker and the texts *"Ready"* and *"Press to start recognizing"*. The user has to press the green button with the microphone to pass to the digitisation activity. In the lower part of the window, two text boxes are presented: the first will show the text recognised by the server and the second its translation. When being pressed, its content will be read aloud in Spanish and English, respectively, using the voice synthesiser of Google. The grey button with the speaker allows the user to listen the audio captured by the device. Apart from that, the user has to his disposal on the application menu the *"Configuration"* and the *"About ..."* buttons. When pressing on the *"Configuration"* button, it goes to the configuration menu, that allows to change two parameters: the IP address and the port of the socket of the server, and the synthesis of voice.

Activity 2: Digitisation of the voice In order to analyse the voice in the server, as it is described in the standard ETSI IS 201 108 [3], this activity digitises the sound that arrives to the microphone of the mobile device with a sampling frequency of $16KHz$, and stores it in an uncompressed audio file. While the voice is getting digitised, the green button becomes red and a text reminds the user that s/he has to *"Talk now"* and to *"Press when finished talking"*. After pressing, it goes to the communication activity.



(a) Establishing the communication by sockets.

(b) Sending the audio file.

(c) Processing on the iAtros server.

(d) Recognition results.

Figure 3. Activity of communication with the server.

Activity 3: Communication with the server This activity establishes the communication by using sockets with the server. It sends to the server the audio file obtained in the previous activity and, while the recogniser is performing the recognition in the server, the client application waits for the response, which is shown to the user.

While the mobile device tries to establish the communication, the text *"Connecting to iAtros server. Please wait ..."* is shown (Figure 3(a)). After the communication is established, this activity sends the audio file and a progress bar is shown (Figure 3(b)). Then, the mobile device waits until the server finishes with the recognition and sends the recognition result, while the text *"Processing on iAtros server. Please wait ..."* is displayed (Figure 3(c)). When the response is received, the recognised Spanish sentence is presented in the first text box and in the lower text box its translation into English (Figure 3(d)). At the same time, it speaks aloud the translation in English using the synthesiser of voice of Google.

Figure 3 shows a complete example of operation of the system. The user said the sentence *"Quiero una habitación tranquila con vistas al mar."*, and the recognition server returned the transcription of the sentence and its translation into English: *"I want a quiet room with a view of the sea."*.

Next, the application goes back from the activity of communication with the server to the main activity, keeping in the text boxes the responses of the server. In this way, the user can read the answer, and even if s/he presses on the text boxes, s/he can listen them until this response is overwritten by a new query of the user. The application reminds the user that admits new queries by the texts: *"Ready"* and *"Press to start recognizing or click the text to replay"*.

2.2 Server

The function of the server is to wait for the requests of the clients and give them a response. The requests of the clients arrive in form of an audio file with the voice of the user and the response that the clients expect is the result of the process of this file by the recogniser: the transcription in Spanish and the translation into English of the sentence said by the user. Our server runs under a GNU/Linux operating system. In this system we installed the speech recognition software iAtros [7,8]. After the server was installed and configured, we developed an application in Java to establish the communication by sockets with the client applications and to execute the automatic speech recognition. The operation is done in four phases:

Phase 1: Listen socket port When the server is started it begins in the this phase. At this stage the server is listening on TCP socket port (e.g., TCP port: 35557) waiting for requests from clients. When a request arrives from a new client, the application sets the communication and the server passes to the audio file reception phase.

Phase 2: Audio file reception In this phase the server application accepts the reception of the audio file. After the reception of the file, the server passes to the ASR phase.

Phase 3: Automatic speech recognition In this phase, the server application uses iAtros to perform the analysis and recognition on the received audio file. Voice analysis encodes the signal file from the mobile device (raw format) into a cepstrum features file. The process is based on the ETSI 201 108 standard, and currently the system uses 23 triangular filters, with a FFT window of 512 samples; each feature vector covers 10 ms., and is formed by 13 cepstrum coefficients with derivative and acceleration. The recognition is performed on the cepstrum features file. The models for the tourist task (acoustic, lexical, and syntactical) were obtained from a previous training process. Acoustic models are Continuous Density Hidden Markov Models, monophone three-state left-to-right topology with loops, with 128 gaussians per state, trained from the Albayzin corpus [9]. Language model is a Finite-State Transducer, that performs the recognition and translation at the same time (as detailed in [1]). After the decoding process, the server passes to phase 4.

Phase 4: Response returns to client applications In this last phase the decoding result obtained by iAtros is sent by the socket to the client application. This response includes the Spanish transcription and its translation into English separated by the symbol |. After sending the response to the client, the server returns to the listening of the port phase, waiting for new client requests.

3 Conclusions and future work

In this demonstration we presented a mobile and distributed speech translation system composed of a server and a client application. The server works under a GNU/Linux operating system in which the iAtros speech recognition software was installed. A Java application was developed to communicate with the client application on mobile devices. The client application was developed for Android version 2.1 and later using the Eclipse development environment, and it is composed of three activities:

- Main: interacts directly with the user, showing the graphical interface with buttons, text boxes and menu options.
- Voice digitisation: records in an uncompressed audio file the signal received at the microphone.
- Communication: sends the audio file and receives the iAtros response to display it in text boxes.

There are some points that can be improved in future versions. On the server part, the extension to other tasks and languages and the addition of a new speaker adaptation module to improve the user experience. And in the case of the client application, the portability to other mobile operating systems and the analysis of the acoustic signal on mobile devices to automatically detect silence and thus, the end of the sentence, making unnecessary to press the button again.

Acknowledgements

This work was partially funded by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and MIT-TRAL (TIN2009-14633-C03-01) projects, by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067, and Vicerrectorado de Investigación of Universitat Politècnica de València under project PAID05-11-0922.

References

1. J. Amengual, A. Castaño, A. Castellanos, V. Jiménez, D. Llorens, A. Marzal, F. Prat, J. Vilar, J. Benedi, F. Casacuberta, *et al.*, "The eutrans spoken language translation system," *Machine Translation*, vol. 15, no. 1, pp. 75–103, 2000.
2. M. Pastor, A. Sanchis, F. Casacuberta, and E. Vidal, "Eutrans: a speech-to-speech translator prototype.," in *Proceedings of EuroSpeech*, pp. 2385–2389, 2001.
3. ETSI, *201 108 v1. 1.2 Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Frontend feature extraction algorithm; Compression algorithms*. 2000.
4. Google, "Android," 2011. http://www.android.com/.
5. J. T. Gironés, *El gran libro de Android*. Madrid: Marcombo, 2011.
6. Google, "Android developers," 2011. http://developer.android.com/index.html.
7. M. Luján-Mares, V. Tamarit, V. Alabau, C. Martínez-Hinarejos, M. Pastor, A. Sanchis, and A. Toselli, "iatros: A speech and handwritting recognition system," *V Jornadas en Tecnologías del Habla (VJTH'2008)*, pp. 75–78, 2008.
8. PRHLT, "iatros," 2011. http://prhlt.iti.es/page/projects/multimodal/idoc/iatros.
9. J. Diaz, A. Rubio, A. Peinado, E. Segarra, N. Prieto, and F. Casacuberta, "Development of task oriented spanish speech corpora," in *Proceedings of EUROSPEECH*, vol. 93, 1993.

COPS: a computer programming tool to cope with functional diversity

Carlos-D. Martínez-Hinarejos[1], Santiago Sánchez-Alepuz[2], and Natividad Prieto-Sáez[3]

[1] Instituto Tecnológico de Informática - Universitat Politècnica de València
Camino Vera s/n, 46022, Valencia, Spain - cmartine@dsic.upv.es
[2] Innovation Betalab - chaumesanchez86@gmail.com
[3] Dep. de Sist. Informáticos y Computación - Universitat Politècnica de València
Camino Vera s/n, 46022, Valencia, Spain - nprieto@dsic.upv.es

Abstract. Computer programming usually requires the use of devices such as screens to see the implemented code and keyboards to input new code. However, these devices are sometimes not suitable for people with functional diversity. In case of sight impairment, tools such as magnification or speech synthesis can be useful to know what is the implemented code. In case of impaired mobility, speech recognition can be useful to input new code. In this work we present an enviroment, called COPS, which is implemented as an Eclipse plug-in, and implements these capabilities for making computer programming feasible for people with functional diversity.

Keywords: Speech recognition, technologies for functional diversity

1 Introduction

Computer programming is an activity that is quite usual for many current jobs. Most people devoted to engineering have to program sometimes, and consequently in most university degrees and professional grades computer programming (at different levels) is a usual subject. The activity of computer programming often uses Integrated Development Environments (IDE) that include editors, compilers, and debuggers.

In any case, IDE tools use the usual channels to interact with the user: screen to show the current code (along with other elements, such as project and packages organisation, menus, etc.), and keyboard and mouse to give inputs (code, menu options via short-cuts or clicks, etc.). When a person with functional diversity faces the activity of computer programming with an IDE, the use of these classical channels may become difficult, uncomfortable, or innapropriate.

For example, the use of the screen as output device (to know what is the current state of the code) is innapropriate for people with severe sight impairment. In some cases, it would be enough to increment font size to make the user read the code, but in other cases other alternatives must be used. One of these possible ways is using speech synthesis (text-to-speech, TTS [1]) to make the

computer play aloud some parts of the code, although other possibilities (e.g., braille touchscreens [2]) could be used.

In the case of code input, the problem appears for people with reduced mobility, specially for those people unable to use their hands to type in with a keyboard or point with a mouse. Thus, an alternative input channel must be defined for these situations. Speech recognition [3] can be used to dictate the code, although other ways (e.g., brain interfaces [4]) could be employed.

In this work we present a tool, developed as a plug-in for a popular IDE, the Eclipse suite [5], that allows the use of font size change, speech synthesis, and speech recognition to ease the computer programming activity for people with functional diversity. The tool is called COPS (of *co*mputer *p*rograming using *s*peech) and implements, by using different libraries and software, these capabilities for programming in the Java programming language in the Eclipse IDE. Section 2 describes the features of the plug-in, and Section 3 discusses its strengths and limitations, along with the proposal of future improvements.

2 COPS functionalities

In this section, the main features of the COPS tool are presented. Subsections 2.1 and 2.2 describe the tools used to improve the code perception by a user with sight impairment, whereas subsection 2.3 describes the use of speech recognition to allow code input for a person with reduced mobility.

The general view of the COPS tool can be seen in the image of Figure 1. In this image, we can notice that there are two text areas: the one on the top is the usual text editing area for the Eclipse IDE, whereas the one on the bottom (corresponding to the tab with the name `COPS plugin view`) mimics the contents of the top area but using black background and white foreground, which allows for more readability. This two text areas are coordinated in such a way that any modification on one of the areas gets reflected in the other text area.

The `COPS plugin view` tab presents four action buttons, that correspond (from left to right) to increasing/reducing font size, perform speech input, and read aloud selected text. Each button calls its corresponding action listener when is clicked, which executes the proper action on the IDE.

2.1 Zooming

When cliking the buttons for font size selection (the first two buttons, with text `A+` and `A-`), the text in the `COPS plugin view` text viewer gets is font size changed. The use of these buttons allows to get the appropriate font size for the user, increasing the readability of the code.

2.2 Reading code: speech synthesis

The speech synthesis of the code requires the previous selection of the code parts to be read aloud. The user can use the mouse or the keyboard to select the corresponding code in the `COPS plugin view` tab, and after that, by cliking

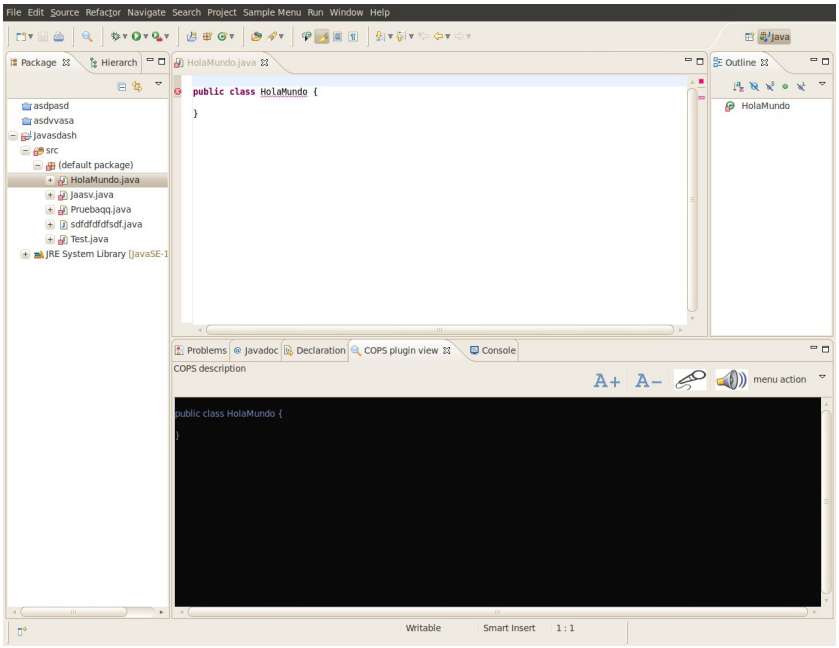


Figure 1. COPS general view.

the speech synthesis button (represented by a speaker icon), the COPS tool will perform the speech syntesis of that segment, that will be played by the computer speaker. Figure 2 shows the detail on how the selected lines appear in the IDE.

Internally, each time the speaker button is clicked, the selected text is synthetised (by default in English) and saved into a file by using the corresponding methods of an instance of the `Voice` class of the package `com.sun.speech.freetts`. This sound file is then played by using an object of the `AudioClip` class, that pertains to the standard `java.applet` package. Currently, the text is synthetised "as-is", with no previous process that makes it more readable for the TTS module.

Figure 2. Text selection in the COPS tool in order to perform text-to-speech.

2.3 Writing code: speech recognition

The recognition button, that is represented by a microphone, is used to activate and inactivate the speech recognition mode. When the button is clicked, the speech recognition switches from the current state to the other. When the tool gets started, the speech recognition is initially disabled.

When the recognition is active and the button is clicked, the method `startRecording` of an object of the `AudioRecorder` class [6] is called; when the recognition is not active, cliking the button produces a call to the methods `stopRecording` and `performRecognition` of the same object. The `AudioRecorded` class employs the classes defined in the `javax.sound. sampled` package to perform the actions. The `AudioRecorder` class that is used in COPS is a slightly modified class with respect to the original class written by Matthias Pfisterer.

The most complex method in the `AudioRecorder` class is the `performRecognition` method; this method converts the audio into the appropriate format, encodes the audio file into a cepstrum data file, and waits for the recognition of the speech recogniser.

The speech recogniser employed in COPS is a modified version of the iATROS off-line recogniser [7]. This modified version waits for a specific cepstrum file to appear in the filesystem, and then performs its recognition, erases the cepstrum file, and writes a file with the result (recognised sentence). This file is detected by the `performRecognition` method, and its content is inserted in the current cursor position in the text area of the `COPS plugin view` tab (which automatically updates the standard text area on the top).

The acoustic models employed in the iATROS recogniser are Continuous Density Hidden Markov Models (CDHMM) with a three-state left-to-right with loops topology, 64 gaussian distributions for state, each of them modelling feature vectors of 12 components plus energy, and their corresponding first and second derivative (39 components in total). These models where trained from the Albayzin speech database [8], since the COPS tool is initially defined to be used by Spanish speakers.

The task vocabulary includes all the Java reserved words, usual identifiers (such as `main`) symbols (such as `.`, `?`, `(`, ...), letters (distinguishing between capital and lowercase letters) and numbers. Lexical models encode all these words into alternative pronunciations which cover different ways of saying a word depending on the English language influence (e.g., `public` can be said `/p u b l i k/` or `/p ae b l i k/`) or reasonable alternatives (e.g., `(` can be said `abrir paréntesis` or `abre paréntesis`).

The language model is a 4-gram with categories. Categories include numbers (including the different formats of real numbers in Java), identifiers (sequences of letters, numbers, and symbols that form a correct Java identifier), strings (starting and ending with `"` and with any character, including spaces, inside) and char constants (starting and ending with `'`, with a single character inside). This category-based structure imposes the use of spelling when dictating identifiers, strings, and so on.

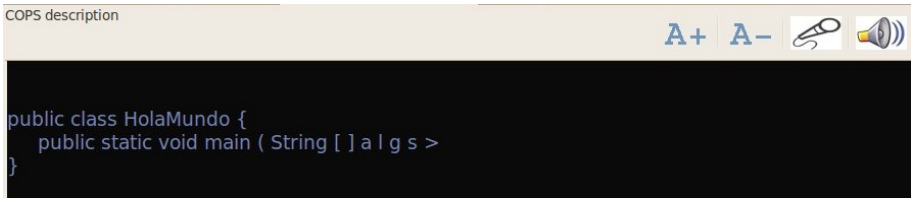


Figure 3. Effect of dictating the code line `public static void main(String [] args)` in the COPS tool.

The 4-gram was inferred from a set of around 300 Java programs, with a total of nearly 36,000 code lines and 387,000 running words, using the SRILM tool. The Java programs were properly tokeninsed and categorised to fit the vocabulary and lexical definitions.

Figure 3 shows the effect of dictating the classical header of a `main` Java method. As can be seen, the speech decoding is not error free, specially on identifiers (`args` is decoded as `algs`), but it is a reasonable first approximation to this problem.

3 Discussion and future work

In this work we presented an interesting application for people with functional diversity that need to program computers. The COPS tool eases the Java code reading and writing by employing enhanced visual and speech interfaces in the Eclipse IDE, although it has not been extensively tested by users and field work must be done in this direction. This tool opens a lot of possibilities for people with functional diversity, that would be able to work with programming environments without assistance of other people.

However, there are still many aspects that can be improved in this tool. At the speech synthesis level, a synthetiser with more quality can be used. At the speech recognition level, the main problem is the spelling of identifiers and strings, which is error-prone and uncomfortable. The spelling could be improved by using a predefined set of identifiers that could be increased by the user (e.g., dictating in spelling mode the identifier and after that choosing a non-spelling form for that identifier to be used during the rest of the program implementation). Another possibility is to include the *a priori* syntax information given by the Java language grammar, that could improve the quality of the recognition of the code lines. Another possible improvement is using speech recognition to manage the Eclipse IDE options (menus, buttons, window selection,...); this could be done by adding an alternative button that recognises orders related to the environment (and not to code dictation). Moreover, the substitution of mouse clicks by special speech orders can make the tool even more useful for people with reduced mobility.

Acknowledgements

This work was partially funded by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and MITTRAL (TIN2009-14633-C03-01) projects, by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067, and by Vicerrectorado de Investigación of Universitat Politècnica de València under projects PAID05-11-0922 and PAID05-11-0947.

References

1. J. Schroeter, "Basic principles of speech synthesis," in *Springer Handbook of Speech Processing* (J. Benesty, M. M. Sondhi, and Y. Huang, eds.), pp. 413–428, Springer, 2008.
2. J. Rantala, R. Raisamo, J. Lylykangas, V. Surakka, J. Raisamo, K. Salminen, T. Pakkanen, and A. Hippula, "Methods for presenting braille characters on a mobile device with a touchscreen and tactile feedback," *IEEE Transactions on Haptics*, vol. 2, pp. 28–39, 2009.
3. L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
4. J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control.," *Clinical neurophysiology*, vol. 113, pp. 767–791, June 2002.
5. J. McAffer and J.-M. Lemieux, *Eclipse Rich Client Platform: Designing, Coding, and Packaging Java(TM) Applications*. Addison-Wesley Professional, 2005.
6. M. Pfisterer, "Audiorecorder.java," 2003. `http://www.jsresources.org/examples/AudioRecorder.java.html`.
7. M. Luján-Mares, V. Tamarit, V. Alabau, C. Martínez-Hinarejos, M. Pastor, A. Sanchis, and A. Toselli, "iatros: A speech and handwritting recognition system," in *V Jornadas en Tecnologías del Habla (VJTH'2008)*, pp. 75–78, 2008.
8. A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *Proceedings of EuroSpeech'93*, (Berlin, Germany), pp. 175–178, sep 1993.

GTMTranscriber: a Subtitling Program for Broadcast News in Galician Language

Antonio Cardenal, Carmen García-Mateo, Alberto Espiña, Paula López, and Laura Docío

Multimedia Technologies Group (GTM)
Universidade de Vigo
E.E. Telecomunicación, 36310 Vigo (Spain)
{cardenal,carmen,aespina,plopez,ldocio}@gts.uvigo.es

Abstract. In this demo we present GTMTranscriber, a program for transcribing and subtitling broadcast news shows of the local Galician television (TVG). It was developed with two main purposes: to demonstrate the capabilities of the recognition system of the GTM group but also as a tool for speeding up the manual subtitling of news shows using automatic speech recognition technology. The system is based on a client-server architecture. The graphical front-end has the ability of replaying any part of the video with embedded subtitles, correcting the subtitles by means of a simple editing console and saving the subtitle file in SRT format. The recognition engine, which runs on a remote server uses diarization and four recognition passes to obtain an almost real-time high quality transcription.

Keywords: speech recognition, subtitles, broadcast news

1 Introduction

GTMTranscriber is an application for subtitling broadcast news shows from the local Galician television (TVG). It is not intended to be a fully automatic subtitling system, but a tool to speed up and automate as far as possible the process of transcribing and subtitling the broadcast news shows. Our main objective was to develop a tool that could be used for subtitling the main news program (Telexornal), before its rerun on the second channel of the TVG, one hour later. A secondary objective was to demonstrate the capabilities of the recognition engine for Galician language of the GTM group.

The antecedents of GTMTranscriber were the projects Transcrigal [5] (in which a database of broadcast news shows was compiled and the basic recognition tools developed), and TACOMA [1]. In project TACOMA we implemented a similar system but oriented to the transcription of generic audiovisual material in Spanish language. During the development of TACOMA we learned several interesting lessons which lead us to rebuild the user interface simplifying the communication layer. The recognition server was also reprogrammed to achieve

real-time operation with minimal accuracy loss.

In the demo session GTMTranscriber will be fully shown, working on a real task. We will transcribe selected news and parts of a real broadcast news program in real time.

2 General overview of the system

GTMTranscriber uses a client-server architecture. It is composed of a graphical front-end (Fig. 1), located on the client side, and a recognition server located on a remote machine. The graphical front-end provides the user with the ability of uploading a video file, starting the recognition process and showing the transcription received from the server. The user may replay any part of the video with the subtitles superimposed just clicking on a segment. By means of a simple console embedded on the main window, the user is also provided with the capability of editing any segment of the transcription and sending it back to the server.

Applying lessons learned from TACOMA, we focused on developing a fast and agile front-end, and on keeping the communication layer between server and client as simple as possible. Following these guidelines, instead of a fully automatic system as TACOMA, we designed GTMTranscriber to be a system where the initiative is always on the client side. The user decides when to check for new transcriptions, when to save any edited subtitles and so on.

With this strategy the communication between client and server is simplified to four commands, which are sent from client to server: to check if the server is busy, to start the recognition process, to stop the recognition process, and to upload a edited subtitle to the server.

For security reasons, the communication is performed via an intermediate Web server. A PHP script is in charge of receiving the audio file and starting the recognition process. Eventually, the recognition server uploads a text file to the Web server indicating the evolution of the decoding process. The front-end allows the user to check the status of the recognizer using this file (Fig. 2). When the transcription of a new section is available, the recognition server puts a SRT (SubRip Text) file on the Web server which will be uploaded and presented by the client on user demand.

3 The graphical front-end

Figure 1 shows the main window of the front-end of GTMTranscriber. It is divided in two main areas: the video playback console, located in the upper area; and the subtitle edition area in the bottom.

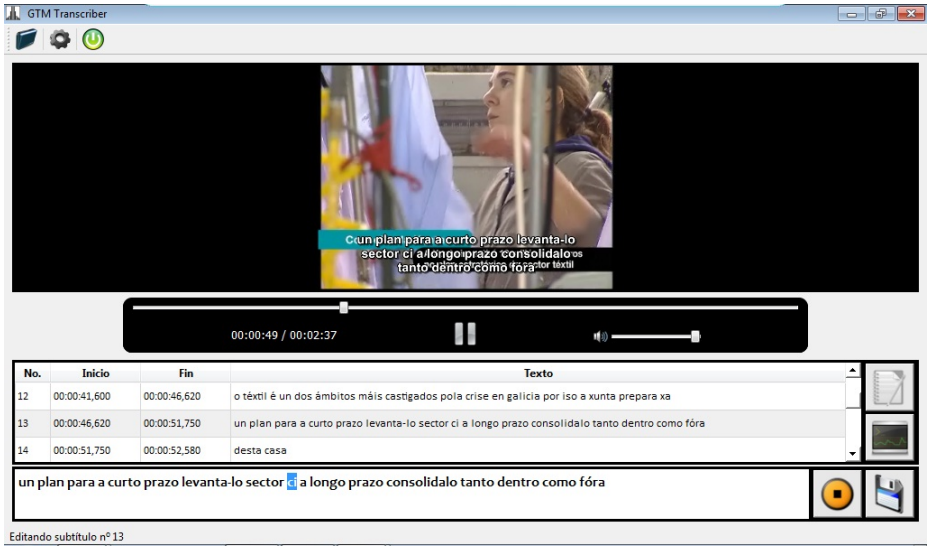


Fig. 1. *Graphical front-end of GTMTranscriber.*

The subtitle console shows the text and time marks of the transcription of each segment. The user may replay any segment with a click of the mouse, or edit the subtitles with a double click. The buttons to the right of the subtitle window lets the user to refresh the transcription, uploading new processed segments, and to inspect the status of the recognition server. The status is presented on a pop-up window (Fig. 2). The two buttons to the right of the subtitle edition console are provided to replay and save the segment under modification

The front-end was programed using Qt language in order to achieve agility and usability. The extraction of the audio part is performed using *ffmpeg*. The audio is downsampled to 16KHz with 16 bits per sample and converted to flac format before sending to the server.

4 Recognition server

The recognition server runs on a remote linux machine. Upon reception, the audio file is parameterized and then segmented using a combined strategy of diarization and VAD (Voice Activity Detection).
The diarization strategy, which is fully described in [3], is implemented in two stages: audio segmentation and classification. First, 12 mel-frequency cepstrum coefficients plus log-energy are extracted from the waveform. Then the audio is segmented by means of a Bayesian information criterion (BIC) based algorithm: anytime a change-point is found by the BIC approach, if its ΔBIC value is lower than a given threshold, this change-point is discarded with a probability

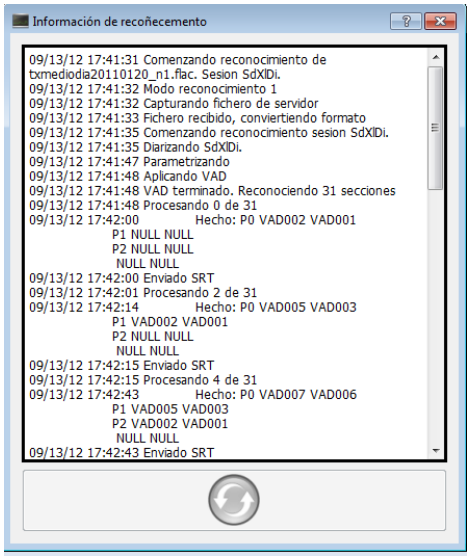


Fig. 2. *Status window in GTMTranscriber.*

that follows a Poisson cumulative density function. The aim of this segmentation strategy is to reduce the number of false alarms.

Once the audio segments are obtained, they are classified into five different classes: speech, music, noise, speech with music in the background, speech with noise in the background and other. Given the feature vectors of an audio segment, a Universal Background Model is maximum a posteriori (MAP) adapted to them in order to represent this segment by means of a Gaussian Mixture Model (GMM). The means of this GMM are concatenated in order to obtain a supervector, which is classified with a support vector machine (SVM) trained for discriminating among the five classes enumerated above.

The segment boundaries and the classification provided by the diarization task are combined with the output of a energy-based VAD to obtain short segments which would be easily managed by the decoder.

The recognition process itself may be performed in up to four different passes, which are applied sequentially to each segment in a pipelined strategy. The four passes are as follows:

1. Acoustic model selection.
2. First full transcription with 3-gram language model (LM).
3. Unsupervised acoustic model adaptation and second full transcription.
4. Reescoring with 5-gram LM.

In the first pass we use eight different acoustic models, three generic: universal, male and female, and five adapted to five of the main speakers of the show. The selection is performed applying a phonetic recognizer over the 10 first seconds of the segment for each HMM (Hidden Markov Model) set, and selecting the set with best acoustic score. The HMMs used are semiphones, with standard MFCC_E_D_A features and CMS extraction.

The second pass uses the HMM set selected in the first pass in conjunction with a 3-gram LM to obtain a first transcription of the segment. The decoder used in this pass was developed in the GTM group and is described in [2].

In the third pass, the automatic transcription is employed to perform a MLLR/MAP unsupervised adaptation of the acoustic models. The new models are used to obtain a new full transcription.

The fourth stage is a 5-gram reescoring pass which is implemented using the SRILM toolkit [7]. We used the tool *segment-nbest* which also provides an hypothesis of the most likely location of segment boundaries. This information is used as a rudimentary punctuation model.

To take advantage of the modern multicore/multiprocessor architecture, the five recognition passes are applied using the parallel processing strategy known as pipelining. The four passes run in parallel, with each pass simultaneously applied to two consecutive segments (this strategy is known as superscalar pipelining [6]). A maximum of eight processes (when the pipeline is full), run simultaneously on the machine. We have found empirically that with this structure the recognition is performed at top speed in a Linux machine with 16 4-core processors Intel Xeon at 2.4GHz.

Additionally it is possible for the user to select a two or three passes scheme sacrificing accuracy for speed.

5 Conclusions and further work

In this paper we have presented GTMTranscriber, a tool for subtitling broadcast news programs in Galician language. The system is capable of transcribing a news program in almost real time, with a short delay imposed by the diarization module.
GTMTranscriber is still under development. There are several features which we would like to include in future versions. To mention a few, a speaker classification module and a punctuation model would be useful additions. A speaker classification module has already been developed on our group, and tested over a broadcast news task. The results may be consulted on [4].

Acknowledgements This work has been supported by the Galician Regional Government (CN2011/019 and 10TIC015E), Spanish Government (TEC2009-14094-C04-04), and the European Regional Development Fund.

References

1. Cardenal, A., Peso, P., Bueno, M., a, A.E., Rodríguez, D.A., Adkinson, L., Pellitero, A.: Tacoma: On-line transcription of audiovisual material. In: FALA2010. pp. 239–242 (2010)
2. Docio-Fernandez, L., Cardenal-Lopez, A., Garcia-Mateo, C.: TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. In: TC-STAR Workshop on Speech-to-Speech Translation. pp. 145–150 (2006)
3. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: A BIC-POISSON-SVM segmentation system for the albayzin'10 audio segmentation evaluation. In: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop. pp. 433–436 (2010)
4. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., Cardenal-Lopez, A.: On the influence of automatic segmentation and clustering in automatic speech recognition. In: IberSpeech2012, accepted for publication on Communications in Computer and Information Science (2012)
5. Mateo, C.G., Tirado, J.D., Fernndez, L.D., Cardenal, A.: Transcrigal: A bilingual system for automatic indexing of broadcast news. In: IV International Conference on Language Resources and Evaluation (2004)
6. Patterson, D.A., Hennessy, J.L.: Computer Organization and Design: The Hardware/Software Interface. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edn. (2007)
7. Stolcke, A.: Srilm - an extensible language modeling toolkit. In: Proceedings International Conference on Spoken Language Processing. pp. 901–904 (2002)

An on-line system adding subtitles and sign language to Spanish audio-visual content

Jordi Porta, Fernando López-Colino, Javier Tejedor, and José Colás

Human Computer Technology Laboratory,
Universidad Autónoma de Madrid, Spain,
jordi.porta@uam.es

Abstract. Deaf people cannot properly access the speech information stored in any kind of recording format (audio, video, etc). We present a system that provides with subtitling and Spanish Sign Language representation capabilities to allow Spanish Deaf population can access to such speech content. The system is composed by a speech recognition module, a machine translation module from Spanish to Spanish Sign Language and a Spanish Sign Language synthesis module. On the deaf person side, a user-friendly interface with subtitle and avatar components allows him/her to access the speech information.

1 Introduction

Spanish Sign Language (LSE) has been considered an official language in Spain since 2007. There are at about one million of Deaf people in Spain and at about 100 000 LSE users. In addition, the ratio between official LSE interpreters and LSE users is 1/221, which falls below the European average of other sign languages.

On the other hand, it has been widely accepted that LSE provides an efficient way for Spanish Deaf people to access to any kind of information, especially speech-based information, no matter the way it is stored (audio or video format) [4–6, 12]. Moreover, the increase in the amount of information stored in a speech-based format makes necessary the quick development of systems and services that allow Deaf people to access to it. To solve this issue, both subtitling and sign language-based systems can be effectively employed [4, 7, 8, 10].

We present an on-line system that produces both subtitles and Spanish Sign Language content aiming at translating the speech content of a video (e.g., TV news, meetings, weather forecast) so that a Spanish Deaf person can access to such content. In so doing, our demo will consist of the subtitled and signed of a Spanish speech content recorded previously in a video format.

Our system takes the speech of the video and sends it towards a speech recognition module that transcribes it into a sequence of words; next, a machine translation module translates this sequence of words into a sequence of glosses; and finally, a Sign Language synthesis module takes this sequence of glosses and generates an animation using an avatar. Both subtitle and avatar components have been included within a user-friendly interface that shows both the subtitles and the LSE content.

2 System description

The system architecture is depicted in Figure 1. The input towards the system is the speech corresponding to the selected video and the output consists of both the word transcription (i.e., subtitles) and the Spanish Sign Language representation of the speech. Our system is divided in the following modules: speech recognition, machine translation and sign language synthesis, which are explained next along with the user interface and middleware.

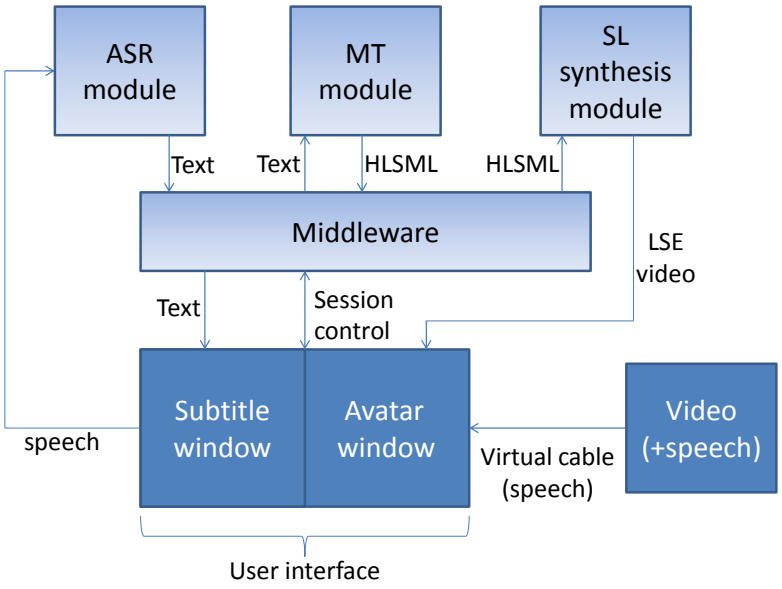


Fig. 1. *The system architecture. 'ASR' stands for Automatic Speech Recognition, 'MT' for Machine Translation, 'SL' for Sign Language and 'HLSML' for High Level Sign Mark-up Language.*

2.1 Speech recognition module

Speech is transcribed by means of the *ATK* software (an HTK-based API for on-line recognition) [14]. As acoustic modeling, and due to the limited training data available, standard 3-state context-independent left-to-right topology Hidden Markov Models were employed for speech decoding. Each state is modeled with 60 Gaussian Mixture Model components. These acoustic models were trained from the ALBAYZIN database [9] for a set of 47 phones [11] in Spanish

plus beginning and end silences with 39-dimensional Mel Frequency Cepstrum Coefficient features. There is also a short pause model that contains a single emitting state and a skip transition. This module also employs a bigram language model during the speech decoding, which was trained from a corpus related to the disability domain, chosen for this demo, and a vocabulary that consists of 2000 words.

2.2 Machine translation module

The machine translation module performs a symbolic translation of a Spanish text into a sequence of glosses representing base signs with morphological annotations. The module has been designed following a transfer-based approach at the level of grammatical functions. The three main phases in which translation can be decomposed are: analysis, transfer and generation. In the analysis phase, the Spanish text is segmented into sentences, words, dates, proper nouns, quantities, etc., using Freeling [1]. Each sentence is parsed using a Spanish grammar with a relatively wide grammatical coverage, which copes with complementation, adjunction, pronominalization, extraction, etc. The type of grammar used is a constraint-based unification grammar with weights expressing preferences. The parser combines a weighted chart with an A* search strategy limited by a beam. In case of failure, the parser applies the shortest-path algorithm in order to recover good partial analyses from the weighted arcs of the chart. The parser produces a constituency tree from which a dependency tree is lift-up. During the transfer phase, the dependency tree is transferred to another LSE dependency tree using a transfer lexicon and some rules for inserting and removing nodes. For example, articles, prepositions and some pronouns are removed, and non-overt subjects are inserted. The transfer lexicon is based on the *Diccionario Normativo de la LSE* [2]. In order to extend the coverage of the lexicon, the morpho-lexical and lexico-semantic relations between words are exploited by using the Wordnet ontology [13]. Finger-spelling is also used in case a Spanish word does not have a sign representation. Generation phase can be subdivided into two subphases. In the first subphase, the dependency tree in LSE is traversed applying linear precedence rules to obtain the surface order of signs. Finally, signs with annotated morphological information are produced in a format according to the input notation (HLSML) of the Sign Language synthesis module.

2.3 Sign Language synthesis module

The Sign Language synthesis module presented in a previous work [3] has been used to synthesize the signs. It receives the message translated by the machine translation module in an HLSML-based format and parses it to produce the sequence of signs in the message. It employs a relational database that stores 500 different signs with their phonetic definition. This phonetic definition is then used to generate the corresponding animation sequence of the whole sentence by means of an avatar. This avatar animation is displayed using a real time rendering API to generate the corresponding video.

2.4 User interface and middleware

A friendly user interface is employed as input/output to/from the system. It sends the speech corresponding to the video to the speech recognition module and gets the video generated by the synthesis module by means of standard RTP connections. The speech is managed through a virtual cable that allows the video display software to send the speech stream to the output side of the cable so that the user interface can receive this stream from the virtual cable input side (as if it was a *real* sound card) and can send it towards the speech recognition module. The subtitle module of the system has been also included in this user interface and hence, this also receives the sequence of words output by the speech recognition module by means of a TCP/IP connection. Therefore, this interface, whose screenshot is depicted in Figure 2, consists of a subtitle window to show the word transcription output by the speech recognition module, and a window to show the signing avatar as main components.

The middleware manages the connection and the dataflow between the modules of the system from standard TCP/IP connections and between the speech recognition module and the subtitle module. Therefore, it sends the speech transcriptions from the speech recognition module to the machine translation module and the subtitle module, and the sequence of glosses from the machine translation module to the Sign Language synthesis module. Thus, the final user (i.e., the deaf person) only needs to install the user interface along with the virtual cable in his/her computer, apart from the appropriate video display software to play the corresponding video.

3 Conclusions

We present an on-line system with subtitling and Spanish Sign Language representation capabilities aiming at representing Spanish speech content. This will allow Spanish Deaf people to access speech information.

Future work will study new application domains such as TV news, weather forecast and so on. This requires extending the vocabulary coverage of the speech recognition and machine translation modules and adding the corresponding signs to the LSE relational database.

References

1. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). pp. 48–55 (2006)
2. Fundación CNSE: Diccionario normativo de la lengua de signos española. Fundación CNSE (2008)
3. López-Colino, F., Colás, J.: Spanish sign language synthesis system. Journal of Visual Languages & Computing 23(3), 121–136 (2012)

Fig. 2. *The user interface and the corresponding video. To avoid cluttering, the subtitle component is not shown in this screenshot.*

4. López-Colino, F., Tejedor, J., Porta, J., Colás, J.: Integration of a spanish-to-LSE machine translation system into an E-learning platform. In: Proceedings of the 6th international conference on Universal access in human-computer interaction: applications and services - Volume Part IV. pp. 567–576 (2011)
5. López-Ludeña, V., San-Segundo, R., de Córdoba, R., Ferreiros, J., Montero, J.M., Pardo, J.M.: Factored translation models for improving a speech into sign language translation system. In: Proceedings of Interspeech. pp. 1605–1608 (2011)
6. López-Ludeña, V., San-Segundo, R., Lutfi, S., Lucas-Cuesta, J., Echevarry, J., Martínez-González, B.: Source language categorization for improving a speech into sign language translation system. In: Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2011). pp. 84–93 (2011)
7. López-Ludeña, V., San-Segundo, R., Martín, R., Sánchez, D., García, A.: Evaluating a speech communication system for deaf people. IEEE Latin America Transactions 9(4), 565–570 (2011)
8. Meinedo, H., Abad, A., Pellegrini, T., Neto, J., Trancoso, I.: The L2F broadcast news speech recognition system. In: Proceedings of FALA. pp. 93–96 (2010)
9. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: Proceedings of Eurospeech. pp. 653–656 (1993)
10. Ortega, A., García, J.E., Miguel, A., Lleida, E.: Real-time live broadcast news subtitling system for spanish. In: Proceedings of Interspeech. pp. 2095–2098 (2009)
11. Quilis, A.: El comentario fonológico y fonético de textos. ARCO/LIBROS, S.A. (1998)
12. San-Segundo, R., Barra, R., D'Haro, L.F., Montero, J.M., Córdoba, R., Ferreiros, J.: A spanish speech to sign language translation system for assisting deaf-mute people. In: Proceedings of Interspeech. pp. 1399–1402 (2006)

13. Vossen, P. (ed.): EuroWordNet. A multilingual database with lexical semantic networks. Kluwer Academic Publishers (1998)
14. Young, S.: ATK: An Application Toolkit for HTK. Engineering Department, Cambridge University (2007)

EntrenaVoz. Videogame feedback system for speech training and rehabilitation

Víctor Osma-Ruiz, Member IEEE, Nicolás Sáenz-Lechón, Member IEEE,
Rubén Fraile, Juan Ignacio Godino-Llorente, Senior Member IEEE

Department of Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Ctra. Valencia, km. 7, 28031, Madrid, España
vosma@ics.upm.es, nicolas.saenz@upm.es, rfraile@ics.upm.es, igodino@ics.upm.es

Abstract. This paper describes EntrenaVoz, a planning application for computer-aided voice biofeedback therapies that allows therapist and patient to have alternative remote sessions to substitute some face-to-face sessions. This is possible thanks to instant communication of the exercises results via Internet.

Keywords. Biofeedback, voice therapy, videogames, EntrenaVoz

1 Introduction

Biofeedback is an essential tool in speech and language training and therapy [1]. A classical scenario is a therapist with a patient, where the former guides the session and informs the patient how to do the exercises and assesses the progress achieved. The therapist usually proposes a work-plan that the patient must comply before the following session.

Computer-aided tools, usually videogames, have been extensively used to produce the required feedback. Traditionally, computer programs have been developed to be used at the therapist office [2][3][4]. They improve the process in two ways: first it is possible to attend several patients at the same time and, second, results are objective. In this case the computer produces immediate feedback to the patient and the therapist can review the recordings and results of the session and schedule more exercises or games [5].

Technological advances in computing, Internet and mobile technologies have made tele-rehabilitation for speech and language disorders a reality. In [6] an on-line system for remote treatment of aphasia is described.

We will concentrate in the area of speech therapy related to phonation skills, that is, breathing, tone, intensity and vocalization. In this area some tools have been integrated in complete speech therapy applications as related in [3][4], where interactive tools are designed to facilitate the acquisition of language skills in the

areas of basic phonatory skills, phonetic articulation and language understanding primarily for children with neuromuscular disorders.

To improve the results and efficacy of the therapy it would be desirable that the therapist could supervise the work of the patient at home and receive the results of the exercises immediately. We present an application that achieves these two goals.

2 Description of the system

The designed system has three different parts or modules that work together following the client-server model. The *management module* allows the therapist to design and to plan exercises that the patient must accomplish by means of the *operation module*; the *data module* is used by the two other modules to store and retrieve information regarding patients and therapies and for communication purposes.

This architecture allows the planning of the exercises and their realization to be performed on different machines. Patients can therefore work at home, communicating with the system through the Internet and freeing the computers at the therapy center, while allowing an immediate analysis of the results by the doctor.

2.1 Management module

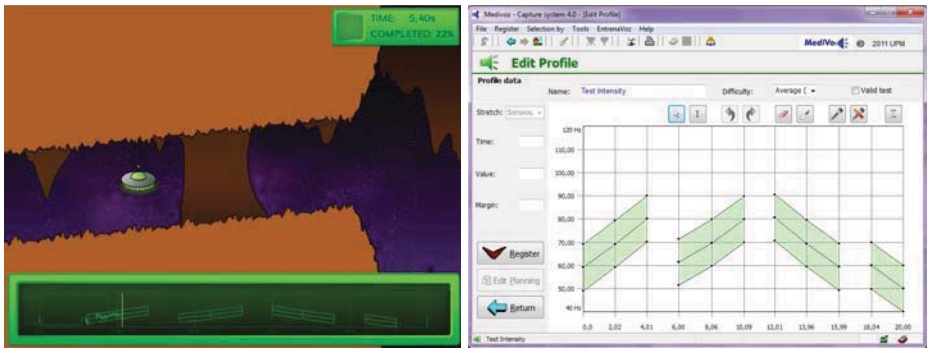
With this module, the therapist can record in a database (based on the data module) different aspects of the clinical history of the patients, as well as design and plan the therapies.

A therapy consists of two types of exercises or videogames: those that are compulsory for the patient and have a strict temporal schedule imposed by the doctor (they must be done once a day, once a week, during a short period of time, or starting on a certain date, etc.) and those that can be performed freely at any time during the therapy.

Currently, the application consists of three different games that can be tuned by the therapist according to the needs of the patients. Each videogame works with a single voice parameter, an so, the current version of EntrenaVoz can work with loudness in a game called *Spaceship*, fundamental frequency in *Glassbreaker* and the differences between voiced and fricative sounds in *Tweet*. The therapist uses these videogames to design compulsory and free exercises.

Training of loudness. Spaceship

This game presents a spaceship flying through a cave (figure 1-a). The patient must control the spaceship, moving it up when he raises the voice and down if he lowers it. The cave moves to the left of the screen with constant speed, and the profile of its walls is designed by the doctor to achieve the desired therapy (figure 1-b). If the loudness is outside the designed profile, the spaceship crashes with the ceiling or the floor of the cave, although it remains undamaged thanks to a protective force shield that allows completing the exercise. The game records the actual contour delineated by the patient.

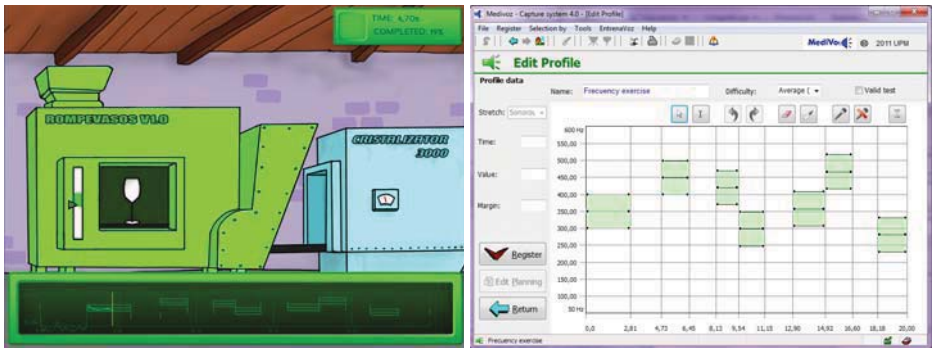


(a) (b)

Fig. 1. (a) Screenshot of the loudness training game. (b) Intensity profile designed by the therapist

Training the fundamental frequency. GlassBreaker

In this game, the patient must keep the fundamental frequency (pitch) within the specified profile (figure 2-b), during a certain time interval. If the patient succeeds in doing so, the glass associated to this interval will break apart, otherwise the glass will escape the machine untouched and a new one will be produced for the next interval.



(a) (b)

Fig. 2. (a) Screenshot of the pitch training game. (b) Fundamental frequency profile designed by the therapist.

Training the differences between voiced-unvoiced sounds. Tweet

In this game the patient must help *Tweet* to get to the other side of his route without getting drenched by the rain nor scratched by the thorny bushes (figure 3-a). These events are determined by the profile designed by the therapist (figure 3-b). When it starts raining, Tweet must generate a voiced sound to open an umbrella, and when bushes appear the chicken must generate a fricative sound to fly past them. As

in the other games, the chicken always gets to the end even if he could not overcome all the obstacles.

(a) (b)

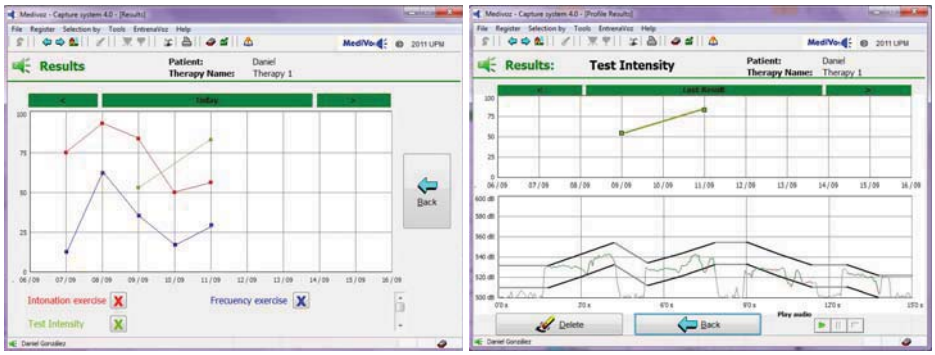
Fig. 3. (a) Screenshot of the game. (b) Profile designed by the doctor

When any of the games finishes, the program shows the success rate, so the patient can have a feedback of the exercise and evaluate his skills. All the therapies and profiles designed by the doctor can be saved in the database, so they can be reused with other patients.

2.2 Operation module

This is the part of the application that patients can carry home to perform the exercises designed for them. It is a simple program that follows these steps:

1. First, the patient must log in the system with a username and password that will be checked against the information stored in the data module to grant him access.
2. After a successful login, all the information regarding the therapies designed for this patient by the doctor will be uploaded from the data module through an intranet or the Internet.
3. The program shows a summary of the therapies the patient must accomplish, with an indication about the type of exercises (compulsory or free exercises) and a comprehensive schedule.
4. From there on, the patient can play the available exercises or view the results of previous ones, where a graph with the evolution of the success rate along time is shown (figure 4-a). The patient can also view the detailed results profile of finished games, as seen in figure 4-b.



(a) (b)

Fig. 4. Displaying of results. (a) Summary graph. (b) Detailed results of a game.

All the results achieved by a patient are stored in the data module, so the therapist can access them immediately within the Management module.

3 Conclusions

Biofeedback is an essential tool in speech and language training and therapy. In this sense, a computer application has been developed, which allows the patient to benefit from this feedback training different aspects of the voice in a challenging way by means of a series of videogames that can be played anywhere. The application also assures the immediate communication of the results to the therapist via the Internet, so the number of face-to-face sessions in the clinic can be reduced.

Acknowledgements

This research was carried out under grants: TEC2009-14123-C04 from the Spanish Ministry of Education and AIB2010DE-00304 from the Spanish Ministry of Science and Innovation. The authors wish to thank Prof. Dr. Katrin Neumann from the University of Bochum (Germany) for her cooperation in the framework of the latter project.

References

1. Maryn Y., De Bodt M., Van Cauwenberge. “Effects of Biofeedback in Phonatory Disorders and Phonatory Performance: A Systematic Literature Review”. Applied Psychophysiology and Biofeedback, vol. 31, No. 1, pp 65-83. 2006.
2. Cagatay M., Ege P., Tokdemir G., Cagiltay N.E. “A serious game for speech disorder children therapy”. Proceedings of the 7th International Symposium on Health Informatics and Bioinformatics, HIBIT2012, pp 18-23. Cappadocia, Turkey, April 2012.

3. Saz O., Yin S.C., Lleida E., Rose R., Vaquero C., Rodríguez W.R. "Tools and Technologies for Computer-Aided Speech and Language Therapy". Speech Communication, vol. 51, pp 948-967. 2009.
4. Rodríguez W.R., Saz O., Lleida E. "A prelingual tool for the education of altered voices". Speech Communication, vol. 54, pp 583-600. 2012.
5. Annema J.H, Verstraete M., Abeele V.V., Desmet S., Geerts D. "Videogames in therapy: a therapist's perspective". Proceedings of the 3rd International Conference on Fun and Games, Fun and Games 2010, pp 94-98. Leuven, Belgium, September 2010.
6. Pompili A., Abad A., Trancoso I., Fonseca J., Martins I.P., Leal G., Farrajota L. "An on-line system for remote treatment of aphasia". Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies, pp 1-10. Edinburgh, 2011.

Interactive and User-Adapted Voice Portals for the City Council*

David Griol, María García-Jiménez, José Manuel Molina, Araceli Sanchis de Miguel

Dep. Computer Science, Carlos III University of Madrid, 28911 - Leganés (Spain)
{david.griol, josemanuel.molina, araceli.sanchis}@uc3m.es,
100025080@alumnos.uc3m.es

Abstract. In this paper we describe a Voice Portal developed to provide municipal information. The different functionalities provided by the system include to query information about the City Council, access city information, carry out several steps and procedures, complete surveys, access the citizen's mailbox to leave messages for suggestions and complaints, and to be transferred to the City Council to be attended by a teleoperator. This way, the voice portal improves the support of public services by increasing their availability, flexibility and control while reducing costs and missed calls.

Keywords: Voice Portals, Spoken Dialog Systems, Speech Interaction, VoiceXML

1 Introduction

Speech Technologies and Language Processing have made possible the development of a number of new applications which are based on spoken dialog systems [1, 2]. One of them are Voice Portals, which facilitate spoken interaction with Internet to provide their users with specific information or web services. These applications can be defined as web portals that can be accessed entirely by means of speech and, usually, DTMF (dual-tone multi-frequency) signaling.

Speech access is then a solution to the shrinking size of mobile devices (both keyboards to provide information and displays to see the results). Besides, speech interfaces facilitate the communication in environments where this access is not possible using traditional input interfaces (e.g., keyboard and mouse). It also facilitates information access for people with visual or motor disabilities.

In addition, Spanish Law on electronic access to public services[1] defines the development of multi-channel access to information as one of the main obligations of the municipalities. This Law also recognizes explicitly the right of citizens to interact electronically with Public Administration. However, a detailed study

* Research funded by projects CICYT TIN2011-28620-C02-01, CICYT TEC 2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485), and DPS2008-07029-C02-02.

[1] http://www.boe.es/aeboe/consultas/bases_datos/doc.php?id=BOE-A-2007-12352

of the current situation in the provision of public services by local councils provides the main conclusion of a lack of electronic public services that meet the functions of providing information, guidance and advice, manage suggestions and complaints, as well as consult official announcements and procedures. Therefore, public voice portals are not only compliant with this Law, but also constitute an efficient tool to provide speech access public services.

2 Interactive Voice Portal to Provide Municipal Information

In this paper we describe a voice portal that integrates different technologies such as the VoiceXML standard, databases, web and speech servers, and several programming languages (SQL, PHP, HTML), which make it more dynamic and flexible and increase its quality, efficiency, and adaptation to the users' specific preferences and needs. The functionalities of the system are to consult information about the City Council (Government Team, Councils, etc.) and the city (history, geographic and demographic data, access to the city, yellow pages, movie show times, news, events, weather, etc.), carry out several steps and procedures (check lists and personal files, book municipal facilities or make an appointment), complete surveys, access the citizen's mailbox to leave messages for suggestions and complaints, and be transferred to the City Council to be attended by a teleoperator.

The *Home module* implements the first dialog that is provided to the user. The options that the system provides to users are divided into five modules that takes into account the type of interaction and data that is facilitated: information, procedures and formalities, surveys, mailbox of the citizen, and human operator. Therefore, this module facilitates the access to the rest of functionalities provided by the portal.

The *Information module* provides specific information about the city. This information has been divided into six categories and classified so that users can easily access each functionality provided by this module:

- *City Council*: This module provides specific information related to the local government, local governing entities, teams and areas.
- *City*: By means of this module it is possible to access information related to the city (history, access, and yellow pages with bars, cafes, restaurants, shops, hostels, hotels and entertainment guides for cinemas and theaters).
- *Thematic Areas*: This module integrates the rest of information that users can consult by means of the voice portal. This information has been divided into 15 areas, which respectively provide general information, competences and contact details. Additional functionalities can be perfectly incorporated by simply adding more static information in the corresponding database of the application.
- *News*: This section provides current news of the municipality, including the date, title and corresponding description of each one of them.

- *Events*: This section reproduces the list of events of the municipality, providing its specific area, title, date, location, and detailed description.
- *Weather information*: Users can get the current weather information from the municipality and a forecast for the next two days.

The *Procedures and Formalities* module provides information about procedures and steps, such as verifying the correct incorporation into specific lists, checking the status of dossiers and documents, booking a municipal facility, or making an appointment related to a specific municipal service. The *Surveys* module allows the City Council to easily and quickly know citizens' opinion about specific issues and queries. Surveys in the application have been designed to be completely anonymous, so the voice portal does not require the users' identification. Once users have finished answering a survey, they are given the options to listen to the answers they have provided.

The *Citizen's mailbox* implements the functionality of recording a user' speech message and store it for further processing. Thus, citizens can provide their requests, complaints, claims or comments at anytime and anywhere. The Citizen's mailbox is then managed by a specific Office of the City Hall. In addition, if users provide their contact information (telephone, mobile phone or email), this Office would contact them to provide a personalized response to their request. Finally, the *Teleoperator* module transfers the user's call to a human operator.

2.1 Architecture of the application

The voice portal has been developed following the client-server paradigm. The architecture is based on two main components: an IVR (Interactive Voice Response) server and a set of web servers. The IVR provides users with pages following the VoiceXML standard[2], the ASR and TTS interfaces, and VoIP and telephony technologies. Different web servers connected to the IVR via Internet provide dialog management facilities, grammars and system prompts, and the access to the information and different web services.

Regarding the VoiceXML server, it receives the users' calls and interprets the documents to provide the required services. The interpreter, in our case the Voxeo Evolution platform[3], also requests the required resources for the application, defines the logic of the services and stores users' session state to interact accordingly. To carry out these actions, the VoiceXML interpreter includes different systems to deal with users' calls, manage the communication with the servers and access the required resources, play audio files, convert text to speech, collect user data, perform voice recording, and manage sessions and events.

Regarding the Web servers, PHP and VXML files are used to implement each service provided by the voice portal, in addition to access MySQL databases containing the specific information. The different functionalities and corresponding files allow users to complete more than one action in each call.

[2] http://www.w3.org/TR/voicexml21/
[3] http://evolution.voxeo.com/

3 Providing a personalized service

The information provided by the voice portal can be classified into static and dynamic information. Static information has been collected from web pages, stored and classified in the databases of the application. Each time users request this information, the system accesses the database and returns this information encapsulated into a VoiceXML file. Examples of this type of information include the history of the city, access information, or contact information of hotels and main offices in the city. Dynamic information includes local news and events, weather information, surveys, and entertainment guides for cinemas and theaters. This information is automatically updated in the application by means of a PHP-based procedure that access the required web pages, carries out a syntax processing of this information, and stores the updated information the database. Each time the user requires this type of information, the system only has to access the database and return it.

All the application dialogs use voice grammars and DTMF, which means that users can access menus by speech or using the phone keys, making the application more accessible. Grammars are encoded following the XML standard format defined by the W3C and, therefore, supported by any VoiceXML platform. In addition, this format allows greater flexibility in terms of grammar structure and debugging. Static grammars deal with information that does not vary over time, including a small number of options to choose from. These grammars are coded in the same file where they are used. Dynamic grammars include information that varies with time and often deal with large amounts of data. These grammars are automatically created using PHP files to manage their contents (creation, obtain contents, modify and update information).

One of the main aspects in the development of the voice portal is the introduction of different functionalities that allow the adaptation of the system taking into account the current state of the dialog as long as the characteristics of each user. On the one hand, we captured the different VoiceXML events considering different messages for the main events: *noinput* (the user does not answer in a certain time interval or it was not sensed by the recognizer), *nomatch* (the input did not match the recognition grammar or was misrecognized) and *help* (the user explicitly asks for help).

Additionally, VoiceXML provides the *property* element to establish the value of a property that affects the behavior of the platform. These properties may be defined for the whole application, for the document, or for a certain element in a form or menu. For the implementation of the voice portal we have tuned the following properties: *Confidencelevel*, *Sensitivity*, *Documentfetchhint* y *Grammarfetchhint*. The property *Confidencelevel* allows to adjust the accuracy of recognition in order to be accepted. The *Sensitivity* allows to adjust the sensitivity of the recognizer. The properties *Documentfetchhint* and *Grammarfetchhint* allow to adjust the usage of the cache to make searches either safer or faster. In our voice portal all these properties are adjusted dynamically depending on the analysis of the generated events and the history of the dialog, following the procedure described in [3].

On the other hand, the voice portal adapts to specific characteristics of the users. It can be used in different languages (Spanish, English, French, German and Italian), as the speech recognition is tuned using the property *xml:lang* and the prompts have been stored in the different languages using different encodings in the database. Also the voice portal stores the telephone numbers from which the users access the system in order to compute which are the most frequent queries and predict the user preferences which can be directly accessed by the user in the next calls in order to safe time and provide a better user experience.

4 Conclusions

This paper describes a voice portal implemented using the VoiceXML standard that constitutes a useful tool to access information of a specific city, carry out procedures and formalities, complete surveys, use a mailbox to make their complaints and requests, and be transferred to a PBX (Private Branch Exchange). The different services offered by the municipal voice portal are divided into modules which are accessed taking into account users decisions during each dialog. These modules are interconnected so that users can complete more than one action.

To our knowledge, there are not voice portals in the Spanish city halls offering the functionalities described. This way, it creates a new communication channel which is useful, efficient, easy to use and accessible. In addition, the voice portal improves the support of public services by increasing the availability, flexibility, control and reducing costs and missed calls. The system is also easily adaptable to the requirements of each municipality. Future works include the development of additional functionalities for the described modules and the adaptation of the voice portal to each user or group of users by taking into account additional languages, previous dialogs and preferred functionalities.

References

1. López-Cózar, R., Araki, M.: Spoken, Multilingual and Multimodal Dialogue Systems. John Wiley & Sons Publishers (2005)
2. Pieraccini, R., Rabiner, L.: The Voice in the Machine: Building Computers that Understand Speech. The MIT Press (2012)
3. Griol, D., Callejas, Z., López-Cózar, R.: Statistical Dialog Management Methodologies for Real Applications. In: Proc. of the 11th SIGdial Meeting. (2010) 124–131

The MovieScript System: Integration of Statistical Methodologies and the VoiceXML standard to develop spoken dialog systems*

David Griol, María Fernández-Vallejo, José Manuel Molina, Araceli Sanchis de Miguel

Dep. Computer Science, Carlos III University of Madrid, 28911 - Leganés (Spain)
{david.griol, josemanuel.molina, araceli.sanchis}@uc3m.es,
100072832@alumnos.uc3m.es

Abstract. In this paper, we present a dialog system developed following a proposal that takes into account the benefits of using standards like VoiceXML and also includes a statistical dialog model to avoid the effort of manually defining the dialog strategy. Our technique is based on the incorporation of a statistical dialog manager that is trained using a labeled dialog corpus. By means of the incorporation of this module, developers only have to define a set of VoiceXML files, each one including a system prompt and the associated grammar to capture users answers for it. This way, the statistical dialog manager automatically decides the following file (i.e. system prompt). The MovieScript system provides information in Spanish about movies, current billboard, and awards in different festivals.

Keywords: Dialog Management, Spoken Dialog Systems, Statistical Methodologies, VoiceXML

1 Introduction

One of the core aspects of developing dialog systems is to design flexible dialog management strategies. The performance of the system depends to a great extent on the quality of this strategy [1]. This design is usually carried out in commercial systems by hand-crafting dialog strategies tightly coupled to the application domain in order to optimize the behavior of the dialog system in that context. This way, these systems are being deployed using industry-wide standards and protocols such as VoiceXML[1], for which different programming environments and tools have been created to help developers. However, the design of the dialog strategy using these methodologies is a very time-consuming process and has the disadvantage of lack of portability and adaptation to new contexts.

* Research funded by projects CICYT TIN2011-28620-C02-01, CICYT TEC 2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485), and DPS2008-07029-C02-02.

[1] http://www.w3.org/TR/voicexml20/

This has motivated the research community to find ways for automating dialog learning by using statistical models trained with real conversations [2, 3]. Statistical approaches can model the variability in user behaviors and allow exploring a wider range of strategies. Although the construction and parameterization of the model depends on expert knowledge of the task, the final objective is to develop dialog systems that have a more robust behavior, better portability, and are easier to adapt to different user profiles or tasks.

We propose to merge statistical approaches with VoiceXML in order to benefit from the flexibility of statistical dialog management and the facilities that VoiceXML offers, which would help to introduce statistical approaches for the development of commercial (and not only accademic) dialog systems. Our technique employs a statistical dialog manager that takes into account the history of the dialog until the current moment in order to decide the next system prompt, whereas the system prompts and the grammars which indicate the possible user responses to them are implemented in VoiceXML. As it is often difficult to acquire and label human-machine corpus which cover an identical domain as the system which is going to be implemented, our approach is also based on the use of simulated users, which is a common practice in machine learning approaches for dialog management.

2 Our proposal to integrate Statistical Dialog Management and VoiceXML

The proposed approach to integrate statistical methodologies in industry applications is based on the automatic learning of the dialog strategy using a statistical dialog management methodology that is detailed in [3]. The selection of the system answer is carried out through a classification process, for which a multilayer perceptron (MLP) is used. To learn the dialog model we propose the use of dialog simulation techniques. Our approach for acquiring a dialog corpus is based on the interaction of a user simulator and a dialog manager simulator [4]. The user simulation simulates the user intention level, that is, the simulator provides concepts and attributes that represent the intention of the user utterance. Therefore, the user simulator carries out the functions of the ASR and NLU modules. Errors and confidence scores are simulated by a specific module in the simulator.

As stated in the introduction, our goal is to combine the flexibility of statistical dialog management with the facilities that VoiceXML offers, which would help to introduce statistical approaches for the development of commercial (and not strictly academic) dialog systems. To this end, our proposal employs the statistical dialog management technique to decide the next system prompt. In addition, the system prompts and the grammars for ASR are implemented in VoiceXML compliant formats, for example, JSGF or SRGS.

Figure 1 shows the architecture designed for the integration of the statistical methodology for dialog management with the functionalities provided by VoiceXML. As can be observed, a VoiceXML-compliant platform (such as Voxeo

Evolution[2]) is used for the creation of Interactive Voice Response (IVR) applications and the provision of telephone access. Static VoiceXML files and grammars can be stored in the voice server. We propose to simplify these files by generating a VoiceXML file for each specific system prompt, as can be observed in the bottom left corner of the figure. Each file contains a reference to a grammar that defines the valid user's inputs for the corresponding system prompt, as observed at the right top corner of the figure.

The dialog system selects the next system prompt (i.e. VoiceXML file) by consulting the probabilities assigned by the statistical dialog manager to each system prompt given the current state of the dialog. This module is stored in an external web server and is implemented using a data structure to store the information that is provided by the user in each dialog turn. The result generated by the statistical dialog manager informs the IVR platform about the most probable system prompt to be selected for the current dialog state. The platform just selects the corresponding VoiceXML file and reproduces it to the user.

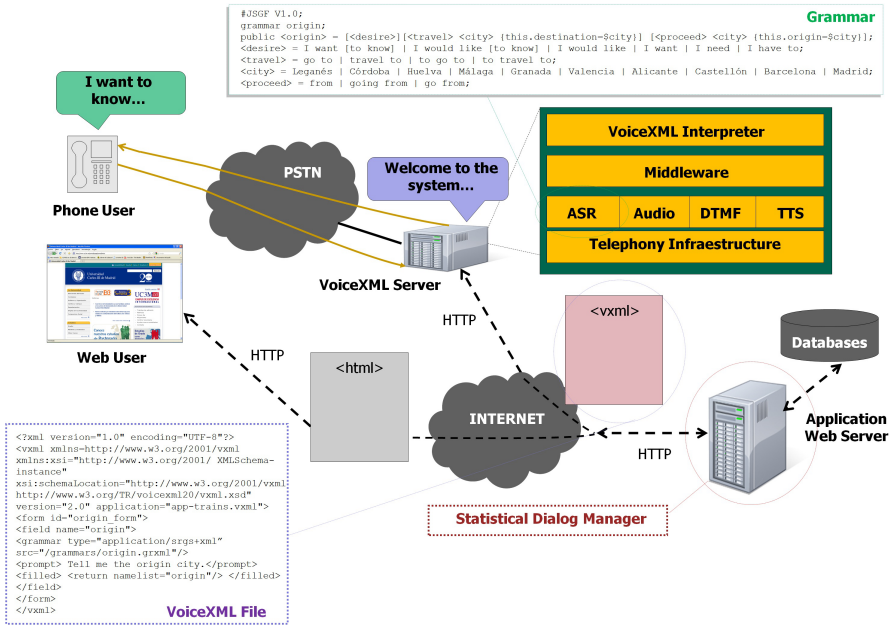


Fig. 1. Architecture for the integration of the proposed statistical methodology for dialog management with VoiceXML

[2] http://evolution.voxeo.com/

3 The MovieScript spoken dialog system

To test our proposal, we have developed a spoken dialog system that provides information in Spanish about movies, current billboard, and awards in different festivals. This information has been extracted from the FilmAffinity movie recommendations website[3].

The system can generate 27 different prompts. These prompts have been automatically generated in VoiceXML using the proposed technique. A total of 5000 successful dialogs was simulated to automatically learn the statistical dialog model. To do this, 53 different scenarios were defined for the simulation considering different objectives to be fulfilled at the end of the dialog that cover the set of functionalities of the system. For example, Figure 2 shows the VoiceXML document to prompt the user about the specific festival in the awarded movies functionality and the corresponding grammar for ASR.

The application is internally divided into three main modules. The first module corresponds to the beginning of the interaction in which the user is welcomed and the system provides detailed instructions about the different functionalities. The second module includes the access to these functionalities, related to information about movies, festivals and current billboard. The third module includes different libraries developed in PHP to access and parse the information extracted from the filmaffinity website and generate the corresponding system prompts.

For the movie information functionality, the user is required to provide the title of the movie. In case that the provided input matches more than one movie, the user is informed about the different options and asked to concrete the required title. Once there is a single matching, the user can consult the synopsis of the film, the cast, the director, and awards.

For the festivals information, users are required to select the name of the specific festival or award (Cannes, Venecia, Berlin, San Sebastian, Toronto, Oscar, Goya, and Golden Globe Awards). Once the system has informed about the name of the awarded movie for the required festival, the user can then obtain detailed information about it, This information is related to the categories previously described for the movie information functionality.

The billboard functionality facilitates the current billboard information for the city provided by the user. This information is updated once a week and users can decide the criteria that is used to present this information (genre, popularity, language, director, or country).

All the application dialogs use voice grammars and DTMF, which means that users can access menus by speech or using the phone keys, making the application more accessible. The system also captures the different VoiceXML events considering different messages for the main events: *noinput* (the user does not answer in a certain time interval or it was not sensed by the recognizer), *nomatch* (the input did not match the recognition grammar or was misrecognized) and *help* (the user explicitly asks for help).

[3] http://www.filmaffinity.com

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml xmlns="http://www.w3.org/2001/vxml"
   xmlns:xsi="http://www.w3.org/2001/
   XMLSchema-instance"
   xsi:schemaLocation="http://www.w3.org/2001/vxml
   http://www.w3.org/TR/voicexml20/vxml.xsd"
   version="2.0" application="app-MovieScript.vxml">
 <form id="awards_form">
  <field name="festival_award">
     <grammar type="application/srgs+xml"
           src="/grammars/festivals.grxml"/>
     <prompt>Tell me the name of the festival or
             award that you want to consult.</prompt>
     <filled>
       <return namelist="festival_award"/>
     </filled>
  </field>
 </form>
</vxml>
```

```
#JSGF V1.0;
grammar festivals;
public <festival_award> = [<desire>]
[<information> <festival> {this.festival=$festival_award}]
[<proceed> <year> {this.edition=$year}];
<desire> = I want [to know | information | information
about] | I would like [to know] | I would like | I want
[information | information about] | I need [information |
information about] | I have to know;
<query> = the winner at | the award at | the awarded film
at | the film that won at | which film won | the name of
the film that won at | last winner at | the awarded movie
at | the movie that won at | which movie won | the name of
the movie that won at;
<festival> = Cannes | Venecia | Berlin | San Sebastian |
Toronto | Oscar | Goya |Golden Globe;
<proceed> = for | at | for the award | for the festival;
<year> = in <one-of> [1929 | 1930 | 1931 | ... | 2012]
</one-of>;
```

Fig. 2. VoiceXML document to prompt the user for the name of a festival or award (left) and grammar to recognize the corresponding value (right)

4 Conclusions

In this paper, we have described a technique for developing dialog systems using a well-known standard like VoiceXML and also including a statistical dialog manager automatically learned from a dialog corpus. The main objective of our work is to reduce the gap between academic and industry perspectives and take the best of both methodologies. On the one hand, the effort that is required for the definition of optimal dialog strategies is reduced. On the other, VoiceXML implementation makes it possible to benefit from the advantages of using the different tools and platforms that are already available to simplify the development of dialog systems. We have applied our technique to develop a complete dialog system that provides information about movies in a very realistic domain. The preliminary objective evaluation that we have already completed with VUI experts at our university has been very useful to verify the correct operation of the system for the described functionalities. We also want to complete a detailed evaluation with real users through quality assessments about users' perception of the system.

References

1. Pieraccini, R., Rabiner, L.: The Voice in the Machine: Building Computers that Understand Speech. The MIT Press (2012)
2. Williams, J., Young, S.: Partially Observable Markov Decision Processes for Spoken Dialog Systems. Computer Speech and Language **21**(2) (2007) 393–422
3. Griol, D., Hurtado, L., Segarra, E., Sanchis, E.: A Statistical Approach to Spoken Dialog Systems Design and Evaluation. Speech Communication **50**(8–9) (2008) 666–682
4. Griol, D., Hurtado, L., Sanchis, E., Segarra, E.: Acquiring and Evaluating a Dialog Corpus through a Dialog Simulation Technique. In: Proc. of the 8th SIGdial Workshop. (2007) 39–42

VITHEA: On-line word naming therapy in Portuguese for aphasic patients exploiting automatic speech recognition

Anna Pompili, Pedro Fialho and Alberto Abad*

L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
{anna.pompili,pedro.fialho,alberto.abad}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/

Abstract. Aphasia is an acquired communication disorder that affects speech and language functionalities at varying degrees. The recovery of lost communication functionalities is possible through frequent and intense speech therapy sessions. The aim of the VITHEA -Virtual Therapist for Aphasia Treatment- project is to exploit speech and language technology (SLT) to facilitate the recovery process of Portuguese aphasic patients, more concretely to recover lost word naming abilities. The proposed system has been designed to behave as a virtual therapist that simulates ordinary speech therapy sessions and, by means of automatic speech recognition (ASR) technology, validates patients performance. In addition to ASR, the system integrates several technological components including animated virtual characters and text-to-speech synthesis (TTS). At the IberSPEECH2012 demo session, the current prototype consisting of two web-based applications will be demonstrated.

Keywords: speech disorder, aphasia, speech recognition, keyword spotting, word naming, virtual therapy, on-line recovery

1 Introduction

Aphasia is a communication disorder caused by the damage of one or more language areas of the brain affecting various speech and language functionalities, including hearing comprehension, speech production, and reading and writing fluency. Among the effects of aphasia, the difficulty to recall words or names is the most common disorder presented by aphasic individuals. In fact, it has been reported in some cases as the only residual deficit after rehabilitation [1]. Several studies about aphasia have demonstrated the positive effect of speech-language therapy activities for the improvement of social communication abilities [2]. Moreover, it has been shown that the frequency and intensity of these therapy sessions positively influences speech and language recovery in aphasic

* This work was partially funded by the Portuguese Foundation for Science and Technology (FCT), through the project VITHEA (RIPD/ADA/109646/2009) and the project PEst-OE/EEI/LA0021/2011.

patients [3]. Typically, word retrieval problems can be treated through semantic exercises like "naming objects" or "naming common actions". The approach usually followed is to subject the patient to a set of stimuli in a variety of tasks.

VITHEA is the first prototype of an on-line platform that incorporates SLT for treatment of Portuguese speakers with aphasia [4]. The system aims at acting as a "virtual therapist", asking the patient to recall the content represented in a photo or picture shown. By means of the use of automatic speech recognition (ASR) technology, the system processes what is said by the patient and decides if it is correct or wrong. Thus, the platform makes available word naming exercises to aphasia patients from their homes at any time, which will allow an increase in the number of training hours, and hopefully a significant improvement in the rehabilitation process.

2 VITHEA: The proposed system

VITHEA is a web-based platform that permits speech therapists to easily create speech therapy exercises that can be later accessed by aphasia patients using a web-browser. During the training sessions, the role of the speech therapist is taken by a "virtual therapist" that presents the exercises and that is able to validate the patients answers. The overall flow of the system can be described as follows: when a therapy session starts, the virtual therapist shows to the patient, one at a time, a series of visual or auditory stimuli. The patient is then required to respond verbally to these stimuli by naming the contents of the object or action that is represented. The utterance produced is recorded, encoded and sent via network to the server side. Here, a web application server receives the audio file encoding the patient's answer and passes it to the ASR system which generates a textual representation of it. This result is then compared with a set of predetermined textual answers (for the given question) in order to verify the correctness of the patient's input. Finally, feedback is sent back to the patient. Figure 1 shows a comprehensive view of this process.

2.1 Automatic Speech Recognition

The ASR module is the backbone of the system: it is responsible for receiving the patient's speech answer and validating the correctness of the utterance for a given therapeutic exercise. Consequently, it strongly determines the usability of the whole therapeutic platform. The targeted task for automatic word naming recognition consists of deciding whether a claimed word W is uttered in a given speech segment S or not. Keyword spotting is an addequate solution to deal with unepected speech effects typical in aphasic patients, such as hesitations, doubts, repetitions, descriptions and other speech disturbing factors. In the current version of the system, acoustic based keyword spotting is applied for word verification. In order to do so, our in-house ASR engine named AUDIMUS [5], that has been previously used for the development of several ASR applications, was modified to incorporate a competing background speech model that is estimated without the need for acoustic model re-training similarly to [6].

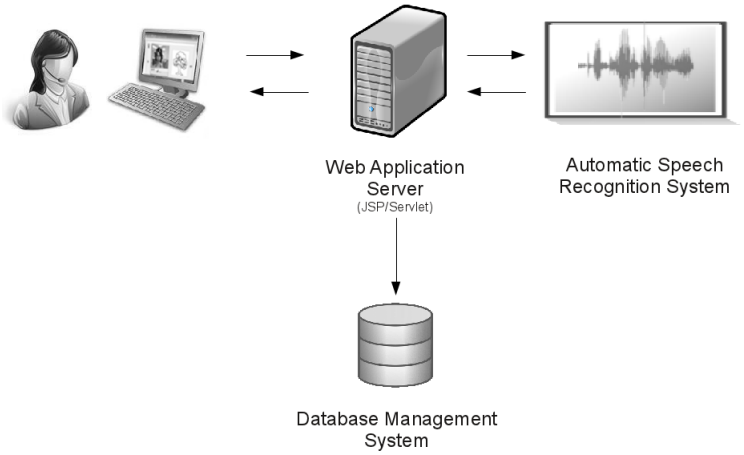


Fig. 1. Comprehensive overview of the VITHEA system.

The baseline speech recognizer AUDIMUS is a hybrid recognizer that follows the connectionist approach [7]. The baseline system combines three MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). The version of AUDIMUS integrated in VITHEA uses an acoustic model trained with 57 hours of downsampled Broadcast News data and 58 hours of mixed fixed-telephone and mobile-telephone data in European Portuguese [8]. The number of context input frames is 13 for the PLP and RASTA networks and 15 for the MSG network. Neural networks are composed by two hidden layers of 1500 units each one. Monophone units are modelled, which results in MLP networks of 39 softmax outputs (38 phonemes + 1 silence). For the word naming detection task, an equally-likely 1-gram language model formed by the possible target keywords and a competing background model is used.

Background speech modelling in a HMM/MLP speech recognizer While keyword models are described by their sequence of phonetic units provided by an automatic grapheme-to-phoneme module, the problem of background speech modelling must be specifically addressed. In this work, the posterior probability of the background unit is estimated based on the posterior probabilities of the other phones. In practice, instead of estimating the posterior, we compute the likelihood of the background speech unit as the mean likelihood of the top-6 most likely classes the phonetic network at each time frame. In this way, there is no need for acoustic network re-training. Further details on this approach and on the complete word verification module can be found in [9].

2.2 Virtual character animation and speech synthesis

The virtual therapist's representation to the user is achieved through a tridimensional (3D) game environment with speech synthesis capabilities. Being a web based platform, the game environment is essentially dedicated to graphical computations (performed locally in the user's computer), while synthesized speech generation occurs in a remote server, thus assuring proper hardware performance. The game environment is based on the Unity[1] game engine and contains a low poly 3D model of a cartoon character with visemes and facial emotions (defined as morph targets), which receives and forwards text (dynamically generated according to the system's flow) to the TTS server. Upon server reply, the character's lips are synchronized with synthesized speech.

Speech synthesis The system integrates our in-house TTS engine named DIXI[10], configured for unit selection synthesis with an open domain cluster voice for European Portuguese. The server uses DIXI to gather SAMPA phonemes[11], their timings and raw audio signal information, which is lossy encoded for usage in the client game. The phoneme timings are essential for a visual output of the synthesized speech, since the difference between consecutive phoneme timings determines the amount of time a viseme should be animated.

Animation Lip synchronization is based on a custom phoneme to viseme mapping, specific to the visemes available in the 3D model. Due to its simplicity (and cartoon nature), this model includes visemes only for *e*, *a* and *o* sounds, therefore it's not suitable for lip reading. Apart from lips, independent idle animations for eyes and head/body were developed, for a more engaging experience.

2.3 The patient and clinician modules

The system comprises two specific web applications: the patient and the clinician modules. They are dedicated respectively to the patients for carrying out the therapy sessions and to the clinicians for the administration of the functionalities related to them.

The patient module The patient exercise interface has been designed to cope with the functionalities needed for automatic word recalling therapy exercises, which includes besides the integration of an animated virtual character (the virtual therapist), text-to-speech synthesized voice [10], image and video displaying, speech recording and playback functionalities, automatic word naming recognition and exercise validation and feed-back prompting. Additionally, the exercise interface has also been designed to maximize simplicity and accessibility. First, because most of the users for whom this application is intended suffered a CVA and they may also have some sort of physical disability. Second, because aphasia is a predominant disorder among elderly people, which are more prone to suffer from visual impairments. Thus, we carefully considered the graphic elements

[1] http://unity3d.com/

chosen, using big icons for representing our interface. Figure 2 illustrates some screen-shots of the patient module.

The clinician module The clinician module is specifically designed to facilitate the management of both patients information and therapy exercises by the therapist. The module is composed of three sub-modules: *User management, Exercise editor, and Patient tracking*. These allow clinicians to manage patient data, to regulate the creation of new stimuli and the alteration of the existing ones, and to monitor user performance in terms of frequency of access to the system and user progress.

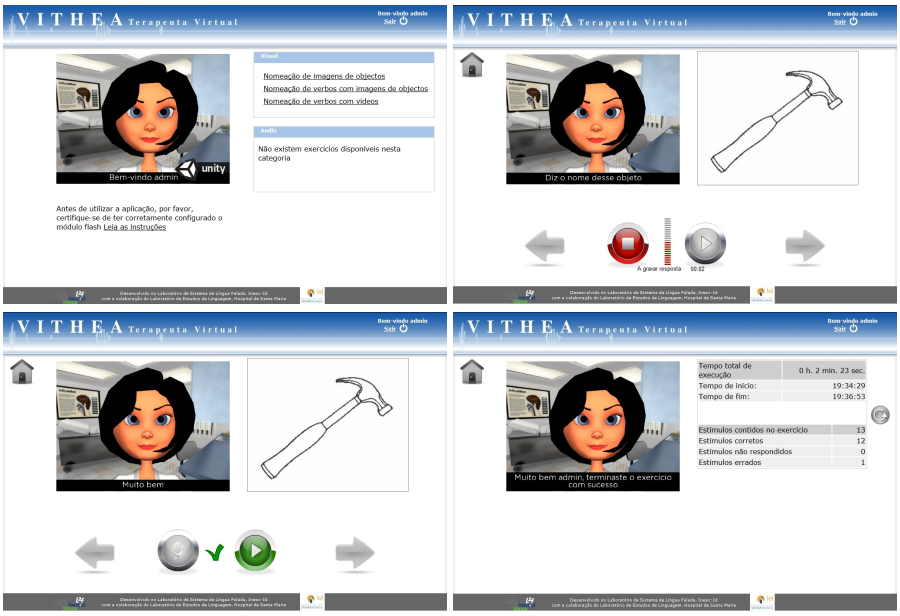


Fig. 2. Screen-shots of the VITHEA patient application.

3 Demo details

During the demo session, the patient [2] and clinician modules [3] will be shown. Since both modules are remote web applications, one or more computers will be provided, although any computer with a built-in or external microphone, Internet connection and a web-browser can be used to access the system. Portuguese speaking attendees will be able to directly test the system: for the patient module they will be able to experience word naming therapy exercises and for the clinician module they will be able to easily create and edit new exercises.

[2] `http://utente.vithea.org`
[3] `http://terapeuta.vithea.org`

References

1. Wilshire, C., Coslett, H..: Aphasia and Language. Theory to practice; chap. Disorders of word retrieval in aphasia theories and potential applications. The Guilford Press, pp. 82–107 (2000)
2. Basso, A.: Prognostic factors in aphasia. Aphasiology. 6(4):337–348 (1992)
3. Bhogal, S., Teasell, R., Speechley, M.: Intensity of aphasia therapy, impact on recovery. Stroke. 34(4):987–993 (2003)
4. Pompili, A., Abad, A., Trancoso, I., Fonseca, J., Martins, I., Leal, G., Farrajota, L.: An on-line system for remote treatment of aphasia. In: Proc. Second Workshop on Speech and Language Processing for Assistive Technologies (2011)
5. Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I., Neto, J.: The L^2F Broadcast News Speech Recognition System. In: Proc. Fala2010 (2010)
6. Pinto, J., Lovitt, A., Hermansky, H.: Exploiting phoneme similarities in hybrid hmm-ann keyword spotting. In: Proc. Interspeech. 2007, pp. 1610–1613 (2007)
7. Morgan, N., Bourlad, H.: An introduction to hybrid HMM/connectionist continuous speech recognition. IEEE Signal Processing Magazine. 12(3):25–42 (1995)
8. Abad, A., Neto, J.: Automatic classification and transcription of telephone speech in radio broadcast data. In: Proc.International Conference on Computational Processing of Portuguese Language (PROPOR) (2008)
9. Abad, A., Pompili, A., Costa, A., Trancoso, I.: Automatic word naming recognition for treatment and assessment of aphasia. In: Proc. Interspeech. 2012 (2012)
10. Paulo, S., Oliveira, L., Mendes, C., Figueira, L., Cassaca, R., Viana, M., et al.: DIXI - A Generic Text-to-Speech System for European Portuguese. In: Proc. International Conference on Computational Processing of Portuguese Language (PROPOR) (2008)
11. Trancoso, I., Viana, C., Barros, M., Caseiro, D., Paulo, S.: From Portuguese to Mirandese: Fast Porting of a Letter-to-Sound Module Using FSTs. In: Proc. International Conference on Computational Processing of Portuguese Language (PROPOR) (2003)

Learning Portuguese with Speech Technologies

Isabel Trancoso[1,2], Thomas Pellegrini[1,3], André Silva[1,2], Rui Correia[1,2], Nuno Mamede[1,2], and Jorge Baptista[1,4]

[1] INESC-ID, Lisbon Portugal
{Isabel.Trancoso,Thomas.Pellegrini,Rui.Correia,Nuno.Mamede}@inesc-id.pt
http://call.l2f.inesc-id.pt
[2] Instituto Superior Técnico, Universidade Técnica de Lisboa
[3] Instituto Superior de Línguas e Administração Campus de Lisboa
[4] Universidade do Algarve, FCHS

Abstract. This set of demos intends to illustrate different applications of speech technologies for Computer-Assisted Language Learning. Although the most typical application in this context seems to be pronunciation training, the emphasis here is on vocabulary learning, and perception. The latter is specially important for European Portuguese as a second language, which is the target of our research. The first demos are aimed at beginners level, and consist of serious games based on 3D, and speech recognition and synthesis technologies, for learning vocabulary and the use of prepositions. The second set of demos are aimed at a higher level, and use recent multimedia documents such as TV broadcast news of the preceding week as training materials. Many speech and language processing technologies are involved in this demo, such as audio segmentation, speech recognition, capitalization, punctuation, topic segmentation and indexation. The generation of the exercises using this data is fully automatic. This may be very valuable for teachers, saving them time in search for motivating materials of appropriate quality, level and topic.

Keywords: CALL, Portuguese, Serious Games

1 Introduction

In the overview of spoken language technology for education [Eskenazi, 2009], the author shows how the use of automatic speech processing for education, and especially for language education, has blossomed in many directions. This paper briefly summarizes our own contributions to this area.

The target language of our CALL (Computer-Assisted Language Learning) systems is European Portuguese (EP). Students of EP as a second language often state that their listening skills cannot cope with spontaneous speech. In fact, one well-known characteristic of EP that distinguishes it from Brazilian Portuguese in particular, is the strong use of vowel reduction and simplification of consonantal clusters, both within words and across word boundaries [Cruz-Ferreira, 2009].

This has been the motivation for adding perception as a major focus in our exercises and games, au par with vocabulary learning and other skills.

The next section describes our 3D serious games, aimed at beginners level, which use speech recognition and synthesis technologies, for learning vocabulary and the use of prepositions. Section 3 describes our CALL applications aimed at a higher level, which use daily adapted multimedia documents such as TV broadcast news as training materials.

2 3D games

The first set of demos aims at teaching vocabulary and also the verbs and prepositions used to describe the spatial relation of objects. Exercises are solved in a game environment making use of a 3D scenario in order to further capture the student's interest, based on the Unity 3D game engine[5].

In the first game, the player controls an avatar through first-person perspective mainly. The scenario consists of an office composed of 5 different rooms, and in each room there are several exercises to be completed. The exercises consist of asking the student to move an object in the scenario to new positions with the use of the mouse, according to a given instruction. For example: *Coloque o objecto A* ***em cima do*** *objecto B* (Put the object A **on top of** the object B). Answers given by the students are automatically evaluated by our game [Silva et al., 2011]. The orders are synthesized using DIXI, our in-house TTS system (Text-to-Speech) [Paulo et al., 2008], as well as the object names, in the tutorial menu. When the player does not position an object in the right place, the game describes the action that was made and the one that should have been made.

The initial feedback we received from students using this game encouraged us to develop a second version of the game in which the student is first shown a fully furnished room whose objects and positions he must memorize. The student must then go to a different room which is totally empty, and refurnish it using exactly the same pieces in the same position using voice input. The student's orders (e.g. put the sofa in the middle of the room) are automatically recognized using AUDIMUS, our in-house ASR system. If the position is correct, the animated object will appear, rotate, and drop in the desired position. A push-to-talk button is used to mark the beginning and end of each utterance.

The third version of the game is only directed at vocabulary coaching. Instead of the office environment, 4 different rooms of a house are used and the recognized vocal input is used to furnish a totally empty room, without specifying positions. Figure 1 illustrates the second version of the game.

3 Daily REAP.PT

The second set of demos involve REAP.PT - the Portuguese version of a vocabulary learning tutoring system, originally developed for American English at

[5] `http://unity3d.com/` (last visited in November 2011)

Fig. 1. 3D game.

Carnegie Mellon University. The system initially focused on vocabulary learning by presenting to students reading material with target vocabulary words in context [Heilman et al., 2006].

The daily REAP.PT version differs from the original system in many aspects, besides the target language: it is based on news that were published during the last 7 days, rather than on documents which were retrieved from the web at a certain time, thus ensuring that they are very recent; it includes both text news and TV news which have been automatically processed; and it allows the students to select a given segment in the text, and listen to the corresponding audio, in a karaoke style. The audio signal is either produced by the TTS system or retrieved from the original video.

In order to be able to use BN videos as learning tools, they need to be automatically segmented, transcribed and indexed. The processing pipeline consists of removing the jingles that usually start and end the news shows, segmenting the audio stream into single-speaker homogeneous speech segments, transcribing the segments automatically with our in-house automatic speech recognition (ASR) system [Neto et al., 2008]. Further modules are then applied to include punctuation, capitalization, and to segment the video into different stories and assign them multiple topic labels. The output of the BN pipeline is comprised of stories with about 300 words each on average. A further filter is applied to automatically estimate the readability level of the stories.

Figure 2 shows a screen-shot of such a page. The left-hand side of the page allows the students to navigate between the different stories of a selected topic, and to see the cooresponding video with captions. On the right-hand side, these same captions are shown with enhanced features. Words with a low confidence estimate are highlighted in red in order to warn the reader that they may be misrecognitions. Target words are highlited in blue, as they will be the focus of the vocabulary exercises the student should complete next.

Fig. 2. Daily REAP.PT.

3.1 Vocabulary perception

Attempting to combine the rich diversity of our BN repository with the motivating aspects of games, we developed "vocabulary perception" exercises, in which the learner transcribes a short video clip by choosing words from lists containing the correct words and some distractors. Our main objective was to give realistic speech for the learners to get used to the pronunciation of native speakers.

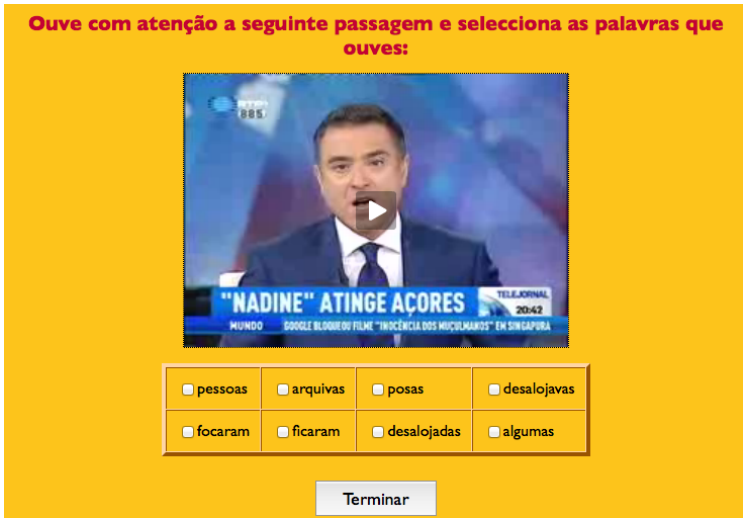


Fig. 3. Tick interface of the vocabulary perception game.

All the exercises are generated in a fully-automatic way. A filtering is needed to discard sentences with probably misrecognized words. A sequence of five filters

was designed to select the sentences: sentence length smaller than 10 words, high ASR confidence measures, syntactic completeness (at least one verb and one common name), large signal-to-noise ratio, descending pitch slope in the sentence boundaries (neutral declaratives).

Different versions of the perception exercise have been tested in order to select the best interface features: audio only or video as well, tick or drag-and-drop interface for selecting the words, with and without slow-down features, etc. Two complementary techniques have also been used to generate the distractors, either based on the confusion networks produced by the recognizer, or on phonetic distances [Pellegrini et al., 2011]. Figure 3 shows one of the interfaces.

4 Conclusions

This paper focused on the CALL demos recently developed at our lab that involve speech technologies. Together with the ones that involve only NLP tools, they are currently under evaluation at different sites: at the University of Algarve and at Cibercursos, a platform for online learning, where the students have teaching sessions run by video conference. The feedback that we have received from the initial surveys is very positive, both from the point of view of learning efficiency, as well as enjoyment of the games and exercises.

Acknowledgments. This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under projects CMU-PT/0005/2007 and CMU-PT/HuMach/0053/2008.

References

[Cruz-Ferreira, 2009] Cruz-Ferreira, M. (2009). European Portuguese. *Journal of International Phonetic Association*, 25:02:90–94.

[Eskenazi, 2009] Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51:832–844.

[Heilman et al., 2006] Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2006). Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension. In *Proc. Interspeech*, pages 829–832, Pittsburgh.

[Neto et al., 2008] Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. (2008). Broadcast news subtitling system in portuguese. In *Proc. ICASSP 2008*, Las Vegas, USA.

[Paulo et al., 2008] Paulo, S., Oliveira, L.-C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., and Moniz, H. (2008). Dixi – a generic text-to-speech system for european portuguese. In *Computational Processing of the Portuguese Language*, volume 5190 of *LNAI*, pages 91–100. Springer.

[Pellegrini et al., 2011] Pellegrini, T., Correia, R., Trancoso, I., Baptista, J., and Mamede, N. (2011). Automatic generation of listening comprehension learning material in European Portuguese. In *Proc. Interspeech*, pages 1629–1632, Florence.

[Silva et al., 2011] Silva, A., Mamede, N., Ferreira, A., Baptista, J., and Fernandes, J. (2011). Towards a serious game for portuguese learning. In *Proc. 2nd International Conference on Serious Games Development and Applications (SGDA 2011)*, volume 6944 of *Lecture Notes in Computer Science*, pages 83–94. Springer-Verlag.

Aromo: Argentine Spanish TTS System

Humberto M. Torres, Jorge A. Gurlekian, and Christian Cossio-Mercado

Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA
Av. Córdoba 2351, 9 Piso Sala 2. Ciudad Autónoma de Buenos Aires, Argentina
hmtorres@conicet.gov.ar,jag@fmed.uba.ar,cgcossio@rucatech.com.ar

Abstract. This paper introduces Aromo text-to-speech system for Argentine Spanish, which was designed for telephony applications and is based on unit selection and concatenation. The system operates as a client-server engine that supports MRCP, SIP and SSML technologies. Perceptual evaluation results show that Aromo's voice achieve high performances in both naturalness and intelligibility.

Keywords: Text-to-Speech; Argentine Spanish TTS; Unit-selection Synthesis

1 Introduction

The Laboratorio de Investigaciones Sensoriales has created Aromo, the first system for converting text into speech designed entirely for the Spanish spoken in Argentina. Its female voice is the first one for this variant of Spanish. It was designed to work as a TTS engine for telephony applications, with a client-server structure that supports MRCP [11], SIP [16] and SSML [17] protocols. It was written in C/C++ under Windows operating system, and works natively in 32 and 64 bits.

2 TTS System Overview

Aromo architecture is modular as other typical TTS systems [3,2,15]. It consists of three main modules: text processing, prosody generation, and unit selection and concatenation. Below, each of the modules will be briefly presented.

2.1 Text Processing

Input text is processed in several sequential stages: first, to put it in orthographic expanded form, by text normalization and tokenization module; then, to extract the sound sequence to be pronounced, by grapheme-to-phoneme conversion module; and finally to extract the text features to be used as input when predicting prosody, by text feature extraction module.

Text normalization covers the following steps: 1) number expansion according to Spanish language rules; 2) translation of abbreviations and acronyms according to a user-defined lexicon or spelling rules; 3) checking of all words in order

to define if they can be uttered by the system and, if not, notification of need of rewriting; and 4) splitting of input text into sentences delimited by punctuation and synthesis control marks.

Grapheme to phoneme conversion module comprises a set of rules that map letters and their corresponding sound according their context [7].

Prosody prediction and unit selection modules require information from the input text as, e.g., sentence type, morphosyntactic structure and location of syllables with lexical stress. Text feature extraction module is responsible for extracting all the features that will be used in the subsequent steps of the process.

Part-of-Speech (POS) tagging is performed using a model trained with a corpus of sentences from newspapers specifically built for this purpose, using a Maximum Entropy tagger [12]. The used tagset [4] is hierarchical —our system uses the first two levels of each tag— and follows EAGLES recommendations [10]. The resulting POS Tagger is available as a web server registered as system service, and its performance level is greater than 98% [19].

2.2 Prosody Prediction

Our system predicts three prosodic features: pauses, segmental duration and fundamental frequency contour (F0).

Pauses are inserted as marked by orthographic symbols or markup tags but also using a prediction model that identify pauses not marked in the text based on POS tags and a set of rules [18].

Duration of each phoneme and pauses are predicted based on the information extracted from the text, using an artificial neural network (ANN) [21] that receives, for example: identity, articulation and sound type of neighboring phonemes; POS tags; lexical stressed condition of syllable; position of syllable into the sentence.

We used the Fujisaki intonation model to predict F0 contour [6], introducing strong linguistic constraints in order to reduce the number of free parameters of the model: one phrase command per intonational phrase located near the beginning of each phrase; and one accent command per content word, located near to the stressed syllable. The parameters of the model are predicted through Classification and Regression Trees [20,22], using the following features as input: location and length of the intonative phrase; identity, location and length of the stressed vowel; identity of the context phonemes; POS tag; distances to the previous and next stressed vowel that has an accent command associated; and parameters of the previous accent or phrase command.

2.3 Unit Selection and Concatenation

Data provided by text processing and prosody prediction modules define a sequence of target phones. With that information the system looks for the optimal sequence of speech units using a generic Viterbi search, minimizing both an objective cost which indicates the distance to the desired unit sequence and a

concatenation cost that takes into account the possible artifacts generated by the paste of the units. In order to quantify these costs, we used physical parameters such as energy, F0 and durations, as well as context parameters such as stress condition, phonetic context, and position within the word and phrase.

In our system we evaluate 37 parameters for calculating total cost. With all these measures, we perform a linear combination that reflects the fitness of each unit to be selected. We decided to use non-linear weight functions so as to avoid generating perceptible artifacts, which greatly degrade the resulting speech quality, and calibrated weight values manually.

Once obtained the optimal sequence of units we proceed to its retrieval from the database and posterior concatenation in order to generate, without modifications, the requested speech.

3 Emilia's Voice Building

The development of a voice requires the adaptation of currently available prosody prediction models and the creation of a database for implementing the unit selection algorithm.

3.1 Building Corpus and Database

When designing textual corpora for a database to be used on a TTS system it is necessary to take into account coverage of phonemes as well as variations in prosody, sentence types and syntactic structures. Our corpus was created taking into account these requisites in a sequential manner. First, we guaranteed the presence of 97% of all Spanish syllables, in both stress conditions and all possible syllabic positions within a word [8]. Then, we added different types of declarative and interrogative sentences with different lengths and syntactic structures. Finally, we checked the phonetic coverage. As a result of this process the corpus resulted on 1851 sentences.

This corpus, of approximately 2.5hs, was recorded by a professional female announcer native of Buenos Aires, instructed to read the sentences with natural tonal variations. Recordings were made in a sound proof chamber, with an AKG dynamic microphone and a sampling of 16Khz/16bit.

The entire corpus was semi-automatically tagged into six tiers: sentences, words, syllables, phonemes, diphonemes and part-of-speech. First we made a first approach by automated methods, and then followed a manual correction done by trained speech therapists. Both phones and diphones transcriptions were automatically aligned with the corresponding speech signal by using two speaker-dependent Hidden Markov Models (HMMs) implemented with HTK speech recognition toolkit [9]. For phonemes we used standard three-state HMMs, and for diphonemes we used four-states model, taking as boundary point the frame in which the transition between the intermediate states occurs.

Sound files and labeling information are used to build the voice database. All information necessary to perform the unit selection process and data for

subsequent concatenation are both created and stored in the data server. This task is done automatically with a tool specially developed for this purpose.

3.2 Building Prosody Models

Although it is not mandatory to adapt prosody prediction models when designing a new voice, it is an recommendable task in order to achieve a good synthesis quality.

To adapt the phone duration models, first we automatically extract the input features from the data corpus and then train an ANN for each phoneme as a predictor of duration, optimizing network parameters configuration for best performance for each phoneme using stratified cross validation. Finally, ANNs are packaged into a file that will be loaded into the data server for future usage.

Similarly, to reestimate the rules that predict the F0 contour model, we use a set of routines that automatically extract the information from the corpus data and then creates a pseudocode with the rules that are to be used for prediction. Finally, we encode and recompile this module.

Prediction pauses model does not need to be updated when creating a new voice.

4 Voice Evaluation

We used three subjective evaluation methods for assessing our TTS system quality, as reported in [5]: ITU P.85 Recommendation, Semantically Unpredictable Sentences (SUS) and Mean Opinion Score (MOS).

ITU.P85 recommendation [14] was created to assess subjective performance for overall quality of speech voice output devices. Topics evaluated are: % of word recognition, overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness —using an ordinal scale of five categories— and acceptance/non-acceptance.

SUS test [1] is focused on evaluating the intelligibility of systems. SUS sentences are syntactically correct but have no meaning and present a very low word predictability. Its measure is the percentage of well recognized words per sentence.

In MOS evaluation [13] listeners use a fixed scale from 1 to 10 to evaluate overall quality. This is a general purpose test and provides average scores for natural and artificial speech.

The stimuli were created according to the recommendations given by three methods. A total of twenty listeners, 23-45 years old, participated in the experiment. Ten of them were Speech Language Therapists (SLT) and other ten were non-expert (non-SLT) without auditory impairments. Figure 1 shows the performance of Emilia voice for the three methods tested and the two groups of listeners.

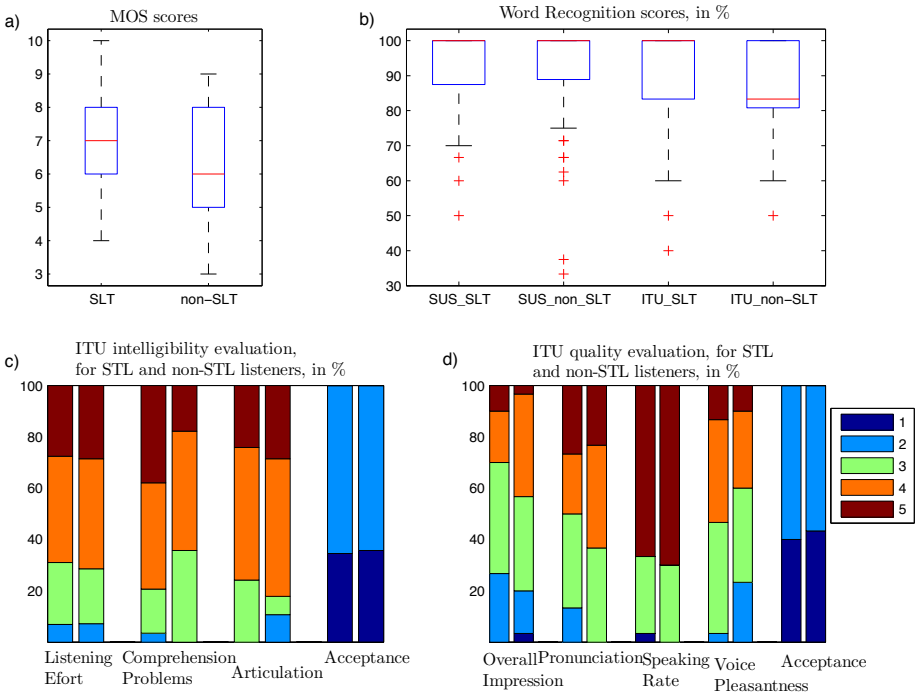


Fig. 1. Perceptual evaluation results: a) MOS quality scores; b) SUS and ITU word recognition scores, in %; c) four topic ITU inteligibility scores, in %; and d) five topic ITU quality scores, in %. The graphs discriminate the results obtained with Speech Language Therapists (SLT) or non-expert (non-SLT). Adapted from [5]

5 Conclusions and Future Work

In this paper we presented a brief description of Aromo TTS system for the Argentine Spanish. We have shown a general outline of its components and how they interact. We also included details of the process of creation of the voice and the results obtained in speech quality perceptual tests. Our system achieved a high performance comparable to other well-known commercial systems [5].

We are now working on creating a new male voice, analogous to the female voice that is currently available. Additionally, we are rebuilding the voice of Emilia in order to obtain greater quality for offline use.

6 Acknowledgements

To the Ministry of Science and Technology and Productive Innovation of Argentina.

References

1. Benoît, C., Grice, M., Hazan, V.: The SUS test. A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication 18(4), 381–392 (June 1966)
2. Clark, R.A.J., Richmond, K., King, S.: Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication 49(4), 317–330 (2007)
3. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Text, Speech and Language Technology, V. 3, Kluwer Academic (1997)
4. EAGLES: `http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html`
5. Gurlekian, J., Cossio-Mercado, C., Torres, H., Vaccari, M.E.: Subjective evaluation of a high quality. In: Proc. of IberSPEECH. p. accepted for publication. Madrid, Spain (November 2012)
6. Gurlekian, J.A., Torres, H.M., Colantoni, L.: Evaluación de las descripciones analítica y perceptual de la entonación de una base de datos de oraciones declarativas de foco amplio para el español hablado en buenos aires. Estudios de Fonética Experimental XIII, 275–302 (2004)
7. Gurlekian, J.A., Colantoni, L., Torres, H.M.: El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados. Fonoaudiológica 47(3), 58–70 (2001)
8. Gurlekian, J.A., Rodriguez, H., Colantoni, L., Torres, H.M.: Development of a prosodic database for an argentine spanish text to speech system. In: Bird, B., Liberman (eds.) Proc. of the IRCS Workshop on Linguistic Databases. pp. 99–104. SIAM, University of Pennsylvania, Philadelphia, USA (December 2001)
9. HTK: `http://www.htk.eng.cam.ac.uk`
10. Leech, G., Wilson, A.: EAGLES recommendations for the morphosyntactic annotation of corpora. Tech. rep., Expert Advisory Group on Lang. Eng. Stds. (1996)
11. MRCP: Media resource control protocol version 2 (mrcpv2) `http://tools.ietf.org/html/draft-ietf-speechsc-mrcpv2-27`
12. OpenNLP: `http://opennlp.apache.org`
13. P.800, I.T.R.: Methods for subjective determination of transmission quality. Tech. rep., ITU (1996)
14. P.85., I.T.R.: Method for subjective performance assessment of the quality of speech voice output devices. Tech. rep., ITU (1994)
15. Schröder, M., Trouvain, J.: The german text-to-speech synthesis system MARY. International Journal of Speech Technology 6, 365–377 (2003)
16. SIP: Session initiation protocol `http://tools.ietf.org/html/rfc3261`
17. SSML: Speech synthesis markup language `http://www.w3.org/TR/speech-synthesis/`
18. Torres, H., Gurlekian, J.: Automatic determination of phrase breaks for argentine spanish. In: Bel, B., Marlien, I. (eds.) Proc. of the Speech Prosody 2004 (SP-2004). pp. 553–556. Nara, Japan (March 2004)
19. Torres, H.M.: Etiquetado de clase de palabras. Tech. rep., LIS (2010)
20. Torres, H.M., Gurlekian, J.A.: Parameter estimation and prediction from text for a superpositional intonation model. In: Proc of the 20 Konferenz Elektronische Sprachsignalverarbeitung. pp. 238–247. TUDpress Verlag (September 2009)
21. Torres, H.M., Gurlekian, J.A.: Argentine spanish segmental duration prediction. In: Proc. of 13th Argentine Symposium on Technology, 41 JAIIO. pp. 156–167. La Plata, Argentina (August 2012)
22. Torres, H.M., Mixdorff, H., Gurlekian, J.A., Pfitzinger, H.: Two new estimation methods for a superpositional intonation model. In: Proc. of INTERSPEECH 2010. pp. 50–53. Makuhari, Japan (September 2010)

Search Engine for Multilingual Audiovisual Contents

José David Pérez[1], Antonio Bonafonte[1], Antonio Cardenal[2], Marta Ruiz[3], José A. R. Fonollosa[1], Asunción Moreno[1], Eva Navas[4], Eduardo R. Banga[2]

[1] *Universitat Politècnica de Catalunya*, [2] *Universidad de Vigo*
[3] *Barcelona Media*, [4] *University of the Basque Country*

Abstract. This paper describes the BUCEADOR search engine, a web server that allows retrieving. multimedia documents (text, audio, video) in different languages. All the documents are translated into the user language and are presented either as text (for instance, subtitles in video documents) or dubbed audio. The user query consist in a sequence of keywords and can be typed or spoken. Multiple Spoken Language Technologies (SLT) servers have been implemented, such as speech recognition, speech machine translation and text-to-speech conversion. The platform can be used in the four Spanish official (Spanish, Basque, Catalan and Galician) and in English.

Keywords: multimedia search, multilingual search, speech recognition, machine translation, speech synthesis, speech to speech translation

1 Introduction

The volume of available information is growing everyday. Information is present in different media, as text, audio, video and images. Furthermore, the information is produced in many languages. In regions with language diversity, as Europe or Spain, even local information exists in several languages.

BUCEADOR is a three years project focused on advanced research in the core Spoken Language Technologies (SLT) such as speech recognition, machine translation, and text-to-speech conversion, and the successful joint integration of all of them in a multilingual and multimodal information retrieval system to access audiovisual contents. The languages considered in the project are Spanish, Catalan, Galician, Basque and English.

This paper describes a showcase developed in BUCEADOR to show the achievements of the project in the above mentioned technologies and their successful joint integration. The platform includes several speech and language engines in different languages. A *multimedia digital library* has been created from sources in all the official languages in Spain. The *documents* include News and debates TV/radio programs, newspapers and magazines.

The remainder of the paper is structured as follows: section 2 presents an overview of the search engine. Section 3 describes briefly the technologies used. Section 4 describes the documents which have been included in the *digital library*.

The user interface and functionality are presented in Section 5. Some examples of use are also presented. The paper ends with a brief summary.

2 Overview of the BUCEADOR search engine

The BUCEADOR search platform retrieves information found in text and audiovisual documents. Many processes are performed off-line, as soon as a new document is introduced in the library. For each system language, a monolingual database is created and indexed. Each database includes the transcription of all the documents in the library. When a new document is added to the library, in case it is either audio or video, the spoken content is transcribed using the speech recognition engines. Then, the original text or the transcribed text is translated into all the languages managed by the system by means of the the translation engines. Each *monolingual* database is indexed so that it can be used by the information retrieval engine. Figure 1 shows a general diagram of the BUCEADOR search platform.

The search engine is a distributed system which includes several web services. The platform interacts with them through Internet. Each partner of the project is responsible for several technologies and languages and provides REST web services to the platform.

In the operative phase, the user selects its own language and introduces a query consisting of a sentence or a set of keywords using either the keyboard (text) or the microphone (speech). If the speech modality is used, a speech recognition service transcribes the input speech into text before doing the search. Then, the information retrieval engine gets the information from the database associated to the user language. The results are presented to the user using either text or speech. If the original information is a text document, the result is also a text document. The user can either read the information or listen to the synthesized speech using the TTS services. If the document is audio or video, the user can also read it or listen to its content. Note that if the language of the audiovisual document is the user language, the user listens to the original audio or reads the transcriptions generated by the speech recognition system. On the other hand, if the language of the document is not the language of the user, then the document has been generated off-line using speech recognition, automatic translation and, optionally, speech synthesis.

As explained above, speech recognition and translation are done off-line, when the documents are introduced, so that the information can be indexed by the information retrieval system. Furthermore, the dubbed version of the audiovisual documents are also generated off-line, when the documents are added to the library. In this way, video files can be produced without any additional delay.

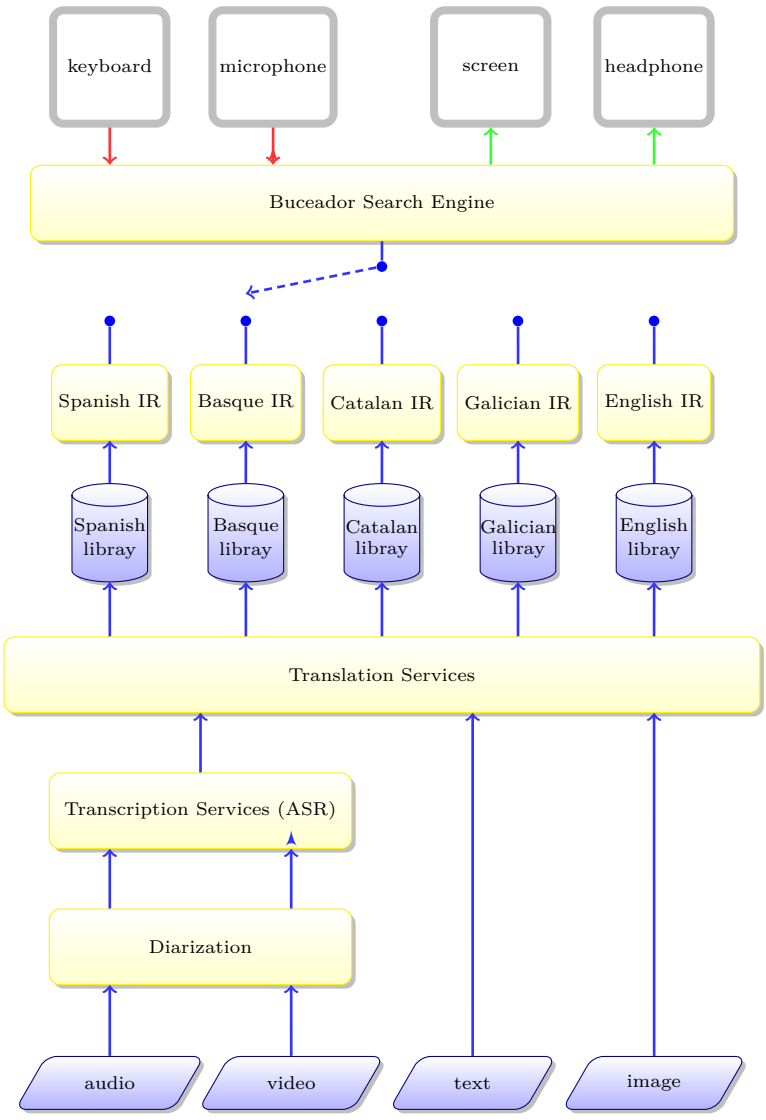


Fig. 1: Overview of the BUCEADOR search engine

3 Technology

This section presents a brief overview about the different technologies involved in this project. For more information, see [1]:

- Speech acquisition: an API programmed in java for the spoken queries. It uses a PHP script on the server side to call that API. The API will send the audio to the recognition servers using a POST method and then the servers will return the query recognized. This recognized query will be passed to the search engine.
- Speech recognition: The speech recognition engine is used in two processes. First, it is used after the diarization process, to transcribe the audio channel in the multimedia documents. It is also used to get the user keywords during the spoken queries.
- Statistical translation: there are translation engines services for Spanish ↔ Catalan, Spanish ↔ Galician and Spanish ↔ English. Spanish is used as the *pivot* language when a direct translation service is not available (for instance, to translate Galician documents into English). Translation from Basque is a very challenging task which is out of the scope of the project.
- Speech synthesis: Once the speech recognition and the translation are done, the speech synthesis engines are used for dubbing the multimedia documents to each possible output language (Spanish, Catalan, English, Galician). Then the dubbed audios are stored along with the original documents. The speech synthesis services are also required in case the user prefers to listen the content of text documents.
- Information retrieval: SOLR [2] has been used for indexation and query retrieval.

4 Multimedia library

The digital library is composed of several resources in all the Spanish official languages (Spanish, Galician, Basque and Catalan).

4.1 The resources

- In the audiovisual part we have:
  The Spanish TC-STAR database [3], is composed of plenary sessions of the European and Spanish parliaments. The speeches from the European Parliament Plenary Sessions (EPPS) have been obtained via Europe by Satellite. It comprises recordings of members and interpreters of the European Parliament speaking in the parliamentary plenary sessions. We have focused only on the Spanish speaking members fo the Parliament. Although in the recordings interventions in Spanish due to the interpreters' work can also be found those parts are not covered by our system. Additionally, the database includes recordings from the Spanish Parliament and the Spanish Congress.

For Galician language we used the Transcrigal-DB which is a database compiled on the University of Vigo during the last years. Transcrigal-DB [4] is composed of recordings of broadcast news programs emitted by the local Galician television (TVG) during years 2002, 2003, 2004, 2010 and 2011. The database contains recordings of 65 news programs of nearly one hour of duration each.
The ETB database is a broadcast news audio database that includes the audio of the news programs of the Basque TV Network (ETB) corresponding to years 2009 and 2010. It is formed by the documents that the journalists prepare and record for a news clip. The audio files contain also interviews and dubbed speech overlapped with the original voice. The files that have associated the corresponding orthographic transcription of the speech and include only one speaker have been selected to be included in the demo. In total there are 1302 files in Spanish and 1276 in standard Basque.
Agora-DB [5] is a Catalan broadcast conversational speech database of the Agora program, that are debates on selected topics from politics, economy or society. Each broadcast follows a repeating format: initially the anchorman presents the current topic, followed by an introduction of invited participants featuring background music. The main part features the debate between the invited participants, usually public figures. During the debate, public opinions are added, either as e-mails or faxes read by the anchorman, or telephone recordings played back featuring background music again. The debate generally comprises spontaneous speech, whereas the introduction of topic and participants features planned speech. Although the main language is Catalan, the recorded broadcasts contain a high proportion of Spanish speaking participants.

– In the text part we have:
The documents from the bilingual newspaper *El Periódico de Catalunya*, included in ELDA catalog, have been added to the BUCEADOR library. This documents are presented in the written form (either in original language or translated). The original sources are in Spanish and Catalan.
Finally the database includes a selection of different sections (*sentencias, consultorio*) of the *EROSKI CONSUMER* magazine. The original sources are in all the official languages of Spain (Spanish, Catalan, Galician, Basque). In these cases, the on-line TTS engines can be used to listen to the content.

4.2 Information organization: XML Documents

When the multimedia database is built, the speech recognition engines [1] are used to transcribe the audio files. The transcriptions are scattered in fragments corresponding to 1 minute of the audio session and stored in XML documents. We will consider that a session corresponds with an audiovisual file (a news program, a TV show, etc.); a session will consist of a set of XML documents.

Usually, speech recognition systems do not produce punctuation. Therefore, one of the problems is the construction of the sentences to be translated and synthesized. From the ASR system we can get the pauses made by the speakers.

In the case of Transcrigal-DB and TC-STAR the solution adopted was to place a punctuation mark when a pause is found. But that solution is not appropriate for Agora-DB spontaneous speech. The sentences produced using this approach would be too short and meaningless. At this moment, this is being investigated: a classifier is being trained using both prosodic and morpho-syntactic features.

When the transcriptions are done, the translation engines [1] translate every transcription to any of the following languages: Spanish, Catalan, Galician, English. As it has been mentioned above, translation from and to Basque is not considered in this project. The translations are stored along with the transcription in the same XML document.

Every XML keeps the most relevant information of the document, see listing 1.1:

- The database and the session to which the document belongs.
- The media type of the session.
- The start/end/duration of every sentence.
- The gender and identification of the speaker.
- The sentence translated into the 4 languages.
- Information about the language and transcription method of every sentence.

Listing 1.1: Xml document

```
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation=
  'http://www.buceador.org/demo/buceador_document.xsd'>
<document_ID> TCSTAR+20040720_1000_1230_ES_SAT+142 </document_ID>
<session_ID> TCSTAR+20040720_1000_1230_ES_SAT </session_ID>
<database_ID> TCSTAR </database_ID>
<URL_media>
   http://www.webs.uvigo.es/gtm_voz/Buceador/TCSTAR/20040720_1000_ES_SAT.mp3
</URL_media>
<media_type>audio/mpeg</media_type>
<segment segment_ID='0' start='8528727' end='8530667'
  duration='1940' start_turn_time='8528727'>
  <speaker id='spk2' gender='unknown'/>
  <original_language>ES_ES</original_language>
  <orth language='ES_ES' method='asr' version='1.0'>
     Muchas gracias señora frassoni.</orth>
  <orth language='GL_ES' method='asr_es2gl' version='1.0'>
     Moitas grazas señora frassoni.</orth>
  <orth language='CA_ES' method='asr_es2ca' version='1.0'>
     Moltes gràcies senyora frassoni.</orth>
  <orth language='EN_UK' method='asr_es2en' version='1.0'>
     Thank you very much, Mrs Frassoni.</orth>
</segment>
<segment segment_ID='1' start='8530667' end='8533307'
  duration='2640' start_turn_time='8528727'>
  <speaker id='spk2' gender='unknown'/>
  <original_language>ES_ES</original_language>
  <orth language='ES_ES' method='asr' version='1.0'>
    Tiene la palabra señor bösch.</orth>
  <orth language='GL_ES' method='asr_es2gl' version='1.0'>
    Ten a palabra señor bösch.</orth>
  <orth language='CA_ES' method='asr_es2ca' version='1.0'>
    Té la paraula senyor bösch.</orth>
  <orth language='EN_UK' method='asr_es2en' version='1.0'>
    You have the floor, Mr bösch.</orth>
</segment>
</document>
```

For each language, the information retrieval indexes are created. According to the user query in a given language the information retrieval system returns the set of documents that match the query. Each result includes the identification of the XML document. This allows presenting the original transcription, the audio or video resource, etc.

5 User Interface

A web page has been designed to integrate the above mentioned technologies. The web page has been programmed in PHP, Javascript and Java.

- The PHP will attend the different web services (query search, subtitles server and synthesizer servers). It also provides access to the information stored in the XML documents.
- The Javascript provides the controls while playing videos, the highlight of the subtitles, the navigation through the results list, the setup of the media player, the selection of the language, the activation/deactivation of the speech recognition. Most of the Javascript code relies on the jQuery library.
- An applet programmed in Java is used to record the speech of a query and send the audio file to a server which will return the keywords of the search to the web server using speech recognition.

Other technologies used in the web page are:

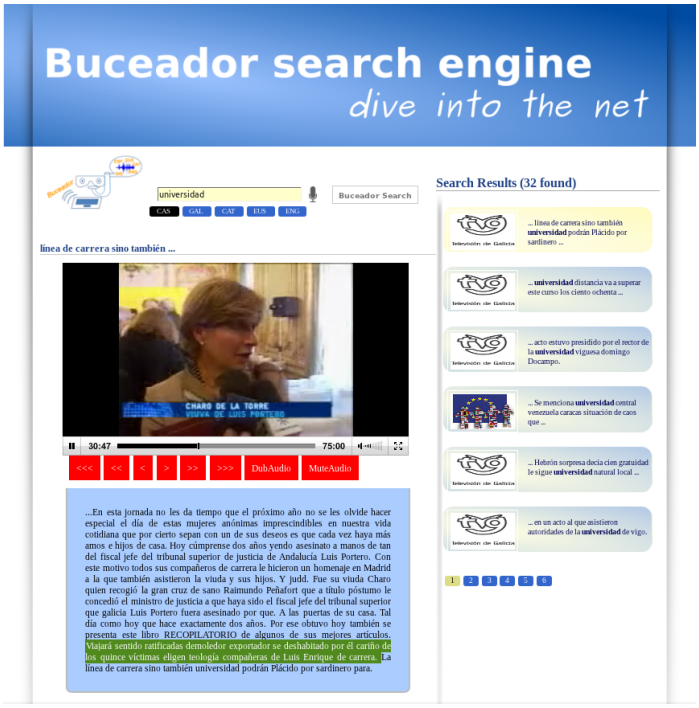
- JW Player: A player that is loaded on the user computer machine responsible of the audio/video playing.
- H264 Streaming module: an apache module for the transmission of H264 video files that enables the user to jump to any part of the video, no matter the length or whether that part has been downloaded yet.
- Rtmp server: the pseudostreaming module is only for H264 video files, so that server is used to transmit audio files.

5.1 Examples

Figure 2a shows a screen capture of the search of the keyword *universidad* (university) in Spanish. There are four main parts:

- The text/audio search field which can be filled either with the keyboard or with the microphone.
- The search results column on the right. Each result can be selected by clicking on the icon.
- The jwplayer part, in this case there is a video playing.
- The subtitles section, with the transcription of the audio. The text highlights are synchronized with the corresponding audio.

Figure 2b is quite similar to the previous figure, but in this case the result is a text document. The user can optionally select part of the text to synthesize it.



(a) Example showing video result

(b) Example showing text result

Fig. 2: Screenshot captures showing the BUCEADOR Search Engine

6 Summary

In this paper, we have presented the showcase of the BUCEADOR project. The goal of this platform is to provide a tool for search on multimedia and multilingual language resources, and present the information in the user language and in the preferred modality (text or audio/video).

One crucial problem is the punctuation of the transcribed audio, in particular in spontaneous (or non-formal) speech. This is essential to produce good translated sentences and intelligible synthetic speech.

7 Acknowledgments

The research leading to these results has been partially funded by the Spanish Government under grant TEC2009-14094-C00. (BUCEADOR project). It has also been supported from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762 (FAUST) and has received funding from the Spanish Ministry of Economy and Competivity through the Juan de la Cierva fellowship program.

References

1. J. Adell, A. Bonafonte, A. Cardenal, M. R. Costa-Jussà, J. A. R. Fonollosa, A. Moreno, E. Navas, and E. R. Banga, "Buceador, a multi-language search engine for digital libraries," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.
2. R. Ku, Apache Solr 3.1 Cookbook. Packt Publishing, 1st ed. edition, July, 2011.
3. H. Van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "Tc-star: New language resources for asr and slt purposes," in *Proceedings LREC*, vol. 2006, 2006, pp. 2570–2573.
4. C. García-Mateo, J. Dieguez, L. Docío, and A. Cardenal, "Transcrigal: A bilingual system for automatic indexing of broadcast news." in *LREC*. European Language Resources Association, 2004.
5. H. Schulz and J. A. R. Fonollosa, "A catalan broadcast conversational speech database," in *I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, Porto Salvo, Portugal, Sep. 2009.

Special Session
Ph.D. Thesis

Application of Speech Technology in the Education of Children's Altered Voice

William R. Rodríguez*

Communication Technology Group, University of Zaragoza,
Maria de Luna 1, 50018 Zaragoza, Spain
wricardo@unizar.es
http://wrodrig.vivolab.es/

Abstract. The thesis addresses the problem of Computer-Aided Voice Therapy by means of a tool called *PreLingua* which provide an interactive voice therapy to a population of individuals with voice disorders. The development of these interactive tool along with the underlying speech technologies that support them requires the existence of speech processing, whose algorithms must be robust with respect to the sources of speech variability that are characteristic of this population of speakers. One of the main problem addressed is how to estimate reliably formant frequencies in high-pitched speech (typical in children and women) and how to normalize these estimations independently of the characteristics of the speakers. This work also presents the result of an experimental study where *PreLingua* was applied in a population with voice disorders and pathologies in special education centers in Spain and Colombia.

Keywords: Altered Voice, Voice Therapy, Speech Processing, Formant Normalization, Vocal Tract Length Estimation

1 Introduction

Working with altered child's voice is a difficult task by speech therapist, even more for those who work with impaired children. Therapist of institutions which offer aid to impaired people and special education schools, despite knowing the benefits of working hand in hand with technology, experience considerable needs and constraints in dealing with altered voice of this kind of population. On the one hand, by the few tools available that have a high acquisition cost, and on the other hand, due most of them are available in other languages than Spanish and other limited options for dealing with vocalic sounds [3], [1], [9], [11]. This thesis addresses this problem by studying the acoustic parameters in voices without alterations using speech technologies (like showed in Figure 1) in order to create free tools in Spanish to allow the education of altered voices in children with or without disabilities. Also, the research included a proposed method to estimate

* Cedesnid - COLCIENCIAS Researcher, Assistant professor of Medicine School and Health Sciences - University of El Rosario. Cl 43 No. 27-25, 111311 Bogotá, Colombia, tel.(057)3133926437.

reliable formant frequencies and how normalize them in order to reduce the high inter-speaker variability which is normal in this kind of voice. After three years of research and development, it was designed a study to apply this tool in real cases of people with disabilities and altered voices in two special education institutes, the quantitative and qualitative results show the benefits and limitations of this tool for training voice as well as other benefits of implementing the proposed technology in pre-linguistic communication and other areas related to special education.

The thesis exposed here can be download from:
`http://i3a.unizar.es/postgrado/descarga_tesis_pdf.php?ver=28`

2 Methodology

The starting point was to know the specialized institutions, their needs, and available tools, and it was also necessary the acquisition of speech corpus in order to estimate its acoustic parameters. Parameters as energy and fundamental frequency or pitch were estimated, also, formant frequencies with difficulty due to the typical high pitch present in this type of voice. This problem was addressed by using traditional techniques in speech processing, and propose a method based on these techniques that allow to estimated formant frequencies in children's speech robustly, and at the same time reducing the high variability by means of a normalization that uses an estimation of the vocal tract length. The development of the tool also required the use of a free graphic motor engine called Allegro [5] in order to link this libraries whit the speech processing algorithms.

3 Speech Tecnhnology in PreLingua

A speech processing diagram block like the one shown in Figure 1 is used for the correct estimation of acoustical parameters, and also show the corresponding game activity in *PreLingua*. In this diagram, after signal preprocessing (DC offset and pre-emphasis), the energy of each frame is calculated (required in *Intensity* activities) and a threshold is applied to determine whether voice from the user is presented on the frame or not (used for *Voice Activity* games). Signal is then windowed using a Hamming window and a LPC [6] analysis is applied to estimate the vocal tract transfer function. From this point, the autocorrelation of the prediction error leads to the extraction of the fundamental frequency value (used in *Tone* activities) and sonority level as the ratio of voiced frames in a certain segment (used in the *Blow* activity). By combining voice activity detection and sonority estimation, it is possible to extract others as the *Vocal Onset* or the *Phonation Time*. *Vocalization* activity and *Articula* make use of formant estimation and normalization. [2], [7], [4], [8], [10]

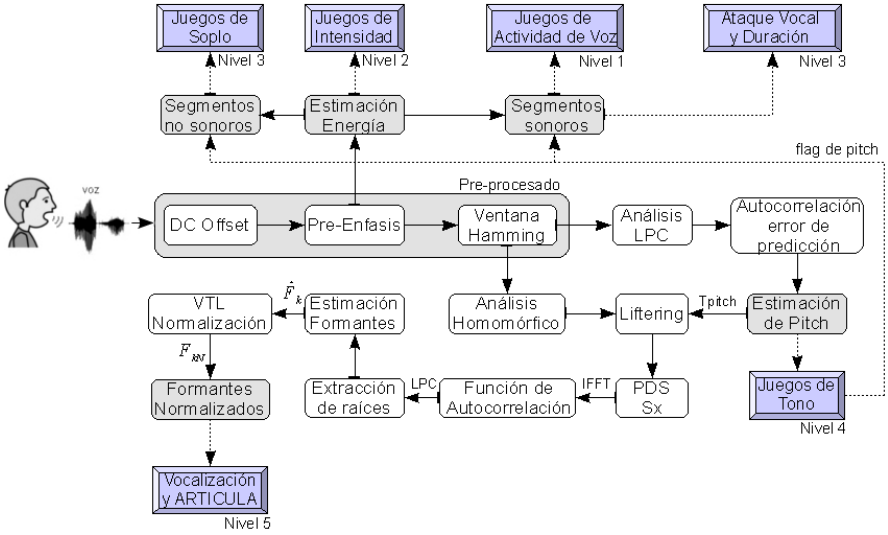


Fig. 1. Block diagram of Speech Technologies implemented in *PreLingua* and its corresponding games.

4 Study

An study was applied on a group of patients from the two educative institutions. 27 subjects were found as potential users of the system and participated in the activities of intensity, blow, tone and vocalic articulation. At the initial stage of the study, each user was evaluated by the therapist through a speech therapy evaluation form in order to assess their voice skills; afterwards, the therapists used *PreLingua* for 12 weeks with each user; and, finally, the therapists repeated the same speech therapy evaluation with each user to assess changes in the voice skills of the subjects. The objective assessment of the subjects was made with data (obtained from the tool) from the first 3 weeks of therapy with *PreLingua* and from the last 3 weeks of therapy with *PreLingua*. Only the MSE between the proposed pattern and the user's performance was considered; and the mean value and standard deviation of this value was calculated for the initial and final sessions separately. The final aim was to detect differences between the results of the activities at the beginning and the end of the therapy to detect possible significant changes in the subjects abilities.

5 Analysis

The observations from the therapists shown by a number of the subjects who participated in the study using *PreLingua* during 12 weeks indicated that the

tool had certainly improved the voice skills of many of them. Elements like blow, intensity and rhythm, which were improved after the therapy by 66.6%, 44.4% and 44.4% of the subjects, respectively, according to the therapists' assessment, were largely influence by the continuous work with *PreLingua*. Therapists observed that *PreLingua* had a high motivational power to attract the attention of the subjects involved in the study (children and adults with mid to severe cognitive disabilities). In general, there was a certain agreement between the tool and the therapists on when a user had improved, although there were a number of subjects with significant improvements in how well they performed different activities which were not assessed by the therapists. However, the comparability between the objective measurement provided by the tool in terms of MSE between a pattern and the subject's production can not be considered to be complete.

References

1. Dr. Speech, http://www.drspeech.com
2. Fant, G.: Acoustic Theory of Speech Production. Mouton & Co., Den Haague, The Netherlands, (1960)
3. Games Program, Key Elemetrics, http://www.kayelemetrics.com/Product Info/3950/3950.htm
4. Necioglu, B., Clements, M., Barnwell, T.: Unsupervised estimation of the human vocal tract length over sentence level utterance. Acoustics, Speech, and Signal Processing, 2000 Vol 3. 1319-1322, (2000)
5. Allegro Graphic Engine, http://www.liballeg.org
6. Rabiner, L., Shafer, R.: Digital Processing of Speech Signals. Prentice-Hall, chapter 4, (1978)
7. Rodríguez, W.-R., Lleida, E.: Formant estimation in children's speech and its application for a spanish speech therapy tool. In: Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE). Wroxall Abbey Estates, United Kingdom, (2009)
8. Shahidur, M., Shimamura, T.: Formant frequency estimation of high-pitched speech by homomorphic prediction. Acoustic Sci. and Tech. 26 (6), 502–510, (2005)
9. Speech Viewer, http://www.synapseadaptive.com/edmark/prod/sv3/
10. Traunmuller, H., Eriksson, A.: A method of measuring formant freqiencies at high fundamental frequencies. In: Proceedings of EuroSpeech'97, vol.1: 477 - 480 (1997)
11. VoxGames, http://www.ctsinformatica.com.br

PhD Dissertation: Advances in Fully-Automatic and Interactive Phrase-based Statistical Machine Translation

Daniel Ortiz-Martínez

D. Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de vera s/n, 46022, Valencia, Spain
dortiz@dsic.upv.es

Abstract. This PhD dissertation, written by Daniel Ortiz-Martínez and supervised by Francisco Casacuberta and Ismael García-Varea, was defended on October 9th 2012 at Universitat Politècnica de València (UPV). The committee members were Marcello Federico (Fondazione Bruno Kessler, Italy), Philipp Koehn (University of Edinburgh, United Kingdom), Hermann Ney (RWTH Aachen University, Germany), Enrique Vidal (UPV) and Juan Miguel Vilar (Universitat Jaume I, Spain). Qualification: "sobresaliente cum laude".
Index Terms: statistical machine translation, interactive statistical machine translation, phrase-based alignment models, incremental learning, online learning

1 Introduction

This thesis explores the area of machine translation (MT). The discipline of MT investigates the use of computer software to translate text or speech from one language to another. Current MT systems use different translation technologies. These translation technologies can be classified into two main approaches: rule-based systems and corpus-based systems. Rule-based systems use a set of translation rules created by human translators to generate their output. Corpus-based systems make use of sets of translation examples (also called corpus or parallel texts) from one language to another.

1.1 Statistical Machine Translation

This thesis approaches MT under the statistical framework. Statistical MT (SMT) systems are a kind of corpus-based MT systems that use parallel texts to estimate the parameters of the statistical models involved in the translation process. Once the statistical models have been estimated, they are used to infer the translation of new sentences. Different statistical translation models have been proposed since the beginning of the research in SMT. The IBM models were the

first statistical translation models used in SMT. In these models, the fundamental unit of translation is a word in a given language. This restricted conception of the translation process does not allow to obtain good translation results due to its inability to capture contextual information. To solve this problem, single words were replaced as the fundamental unit of translation by multiple words to originate a new family of statistical translation models. Among the different multi-word statistical translation models that have been proposed so far, the so-called phrase-based models currently constitute the most commonly used translation models in SMT. In this context, a phrase consists in a set of consecutive words of the source or the target sentence. More recently, state-of-the-art SMT systems also use the so-called log-linear models, which are composed of a set of scoring functions, each one with its corresponding weight (such weights can be viewed as a measure of how important the associated component is with the translation process). The log-linear models used in SMT are strongly focused on the use of phrase-based models as scoring components.

The building process of an SMT system involves addressing three problems:

1. the *modelling problem*, that is, how to structure the dependencies of the statistical model that relates source and target language sentences.
2. the *training problem*, that is, how to estimate the model parameters given the training data.
3. the *search problem*, that is, how to find the best translation candidate among all possible target language sentences.

1.2 Interactive Machine Translation

Current MT systems are not able to produce ready-to-use texts. Indeed, MT systems usually require human post-editing in order to achieve high-quality translations. This motivates an alternative application of MT in which the MT system collaborates with the user to generate the output translations. This alternative application receives the name of computer-assisted translation (CAT). In this thesis, we will focus on a specific instantiation of the CAT paradigm which receives the name of interactive machine translation (IMT). In the IMT framework, the user obtains her desired translations in a series of interactions with a fully-fledged SMT system. Specifically, the system initially generates a translation without human intervention and after that, the user validates a prefix of the translation and introduce the next correct character of it. With this information, the IMT system returns the suffix which best completes the prefix given by the user. This process is repeated in a series of interactions until the user gets the sentence she has in mind.

2 Thesis Summary

This thesis presents different contributions in the fields of fully-automatic statistical machine translation (SMT) and interactive statistical machine translation (IMT). The proposed techniques are focused on the use of phrase-based models, which constitute the state-of-the-art in SMT.

2.1 Scientific and Technologic Goals

The scientific and technologic goals of this thesis can be divided into two groups, fully automatic and interactive phrase-based SMT goals:

1. **Fully automatic phrase-based SMT**:
   - With respect to the modelling problem, propose an alternative phrase-based model derivation that allows to obtain a set of probabilistic models governing different aspects of the translation process, allowing to formally justify the use of scoring functions that are commonly found in standard translation tools.
   - Regarding the estimation problem, describe a new estimation procedure that reduces the strong heuristic component of the standard estimation algorithm while maintaining feasible computational costs.
   - With respect to the search problem, study the capabilities of different search algorithms than can be found in the literature. In addition to this, study the problem of generating alignments at phrase level. The generation of phrase-based alignments has not been extensively studied in the literature and it is an interesting problem since a range of different applications from phrase-based SMT systems to machine-aided NLP tools can benefit from the availability of phrase-based alignments.
   - Regarding both the training and the search problems, study techniques to deal with very large corpora during both the estimation and decoding stages. This also constitutes an interesting problem, since the huge size of current training corpora can make both estimation and decoding unfeasible.
2. **Interactive phrase-based SMT**:
   - Define new techniques to tackle with one of the main problems that arise in IMT, specifically when the user sets a prefix which cannot be explained by the statistical models used by the IMT system.
   - Define an IMT system which is able to extend the statistical models in real time when a new translation is validated by the user. Current state-of-the-art IMT systems are not able to take advantage of user feedback.

2.2 Thesis Achievements

The list of achievements of this thesis with respect to the scientific and technologic goals mentioned above is as follows:

Fully-Automatic Phrase-Based SMT Achievements

- **Specific phrase-based model derivation**: The proposed derivation allowed us to formally justify the set of components used in a state-of-the-art loglinear model. Its performance in terms of translation quality was comparable with that obtained by a well-known translation toolkit (the Moses toolkit) for three different corpora.

- **Improved phrase-based model estimation**: The proposed alternative estimation technique effectively reduces the heuristic component of the standard estimation technique and can be efficiently executed. The likelihood of the resulting models was improved but results not (one possible explanation is that the algorithm is overfitting the probability distributions). We think that the proposed estimation can be used to obtain better models.
- **Phrase-based model estimation from very large corpora**: The proposed technique allows us to train corpora of an arbitrary size. The technique is easilly parallelizable and constitutes an application of a well-known software engineering framework.
- **Branch-and-bound search for phrase-based SMT**: We obtained a branch-and-bound search algorithm that can be parameterized to obtain different algorithms (best- and breadth-first, single- and multiple-stack algorithms, etc.). Additionally, a theoretical and empirical study of the properties of such algorithms was carried out.
- **Efficient decoding with large phrase-based models**: A technique inspired in the concept of cache memory to decode with large phrase-based models was proposed. The proposal successfully transforms memory requirements into hard disk requirements, introducing a very low time overhead.
- **Generation of phrase-based alignments**: A technique to generate phrase alignments was proposed. Such technique is based on smoothing techniques and a specific search algorithm.
- **Development of open-source software**: The open-source THOT toolkit for statistical machine translation was released. It can be downloaded from SourceForge.

Interactive Phrase-Based SMT Achievements

- **Alternative IMT techniques**: IMT systems based on partial phrase-based alignments and on error-correction techniques were proposed. IMT based on partial phrase-based alignments consitutes an application of the phrase alignment generation techniques proposed in this thesis. The technique based on error-correction techniques succesfully introduces stochastic error-correction models in the statistical formalisation of the IMT process. The proposed systems significantly outperformed the results obtained by state-of-the-art IMT systems. Both techniques were implemented in an IMT prototype
- **Online learning for IMT**: an IMT system based on an incrementally updateable SMT model was described. The proposed system is able to learn from scratch of from previously estimated models. As far as we know, this is the first complete system of its kind reported in the literature. Significant improvements in the results were obtained with respect to conventional systems. The proposed online learning techniques were applied in an IMT scenario but they can also be applied without modification in fully-automatic machine translation tasks, making the proposal of potential interest in real-world and commercial applications.

3 Selected Contributions

In this section we briefly describe the main contributions presented in the thesis.

3.1 Generation of Phrase-based Alignments

The generation of phrase-level alignments is not a trivial task due to problems with unseen events, which may cause that a given sentence pair cannot be adequately aligned. The phrase-level alignment generation techniques that were developed during the thesis allows us to solve this problem and can be useful in a wide range of applications, including multi-source SMT, Viterbi-like estimation of phrase-based models, discriminative training and training of phrase segmentation models just to name a few.

The problem of finding the best alignment at phrase level has not been extensively addressed in the literature. A first attempt can be found in [4], where different techniques to obtain alignments at phrase level are proposed. However, the proposed techniques heavily rely on word alignment models or on word alignment matrices.

A very straightforward technique can be proposed for finding the best phrase-alignment of a sentence pair. This can be conceived as a sort of *constrained* translation. In this way, the search process only requires the use of a regular SMT system which filters its phrase table in order to obtain those target translations for the source sentence that are compatible with the given target sentence.

In spite of its simplicity, this technique has no practical interest when applied on regular tasks. Specifically, the technique is not applicable when the alignments cannot be generated due to coverage problems of the phrase-based alignment model (i.e. one or more phrase pairs required to compose a given alignment have not been seen during the training process).

One possible way to overcome the above-mentioned coverage problems requires the definition of an alternative technique that is able to consider every source phrase of the source sentence as a possible translation of every target phrase of the target sentence. Specifically, our proposal is based on a search algorithm that enables efficient exploration of the set of possible phrase-alignments for a sentence pair. The search process is guided by a log-linear model which includes smoothed phrase-based models that allows us to overcome coverage problems as well as specific componentes specifically designed to improve the phrase alignments.

Experimental results for a well-known shared task on word alignment evaluation were obtained. The results showed the great impact of the smoothing techniques on alignment quality. Additionally, the proposed alignment generation algorithm was easily adapted for its use in an alternative IMT technique also presented in the thesis.

3.2 IMT based on Stochastic Error Correction Models

As was mentioned in section 2.1, a common problem in IMT arises when the user sets a prefix which cannot be explained by the statistical models. This problem

requires the introduction of specific techniques in the IMT systems to guarantee that the suffixes can be generated. The vast majority of the IMT systems described in the literature use error-correcting techniques based on the concept of edit distance to solve the coverage problems. Such error correcting techniques are applied on a word graph data structure that represents possible translations of the given source sentence. The applied error-correcting techniques, although they are not included in the statistical formulation of the IMT process, are crucial to ensure that the suffixes completing the prefixes given by the user can be generated.

An alternative formalisation that introduces stochastic error-correction models in the statistical formulation of the IMT process was proposed. The proposed technique generates the suffixes required in IMT by partially aligning a prefix of the target hypotheses with the user prefix. Once the partial alignment is determined, the suffix is given by the unaligned portion of the target sentence. In our proposal, stochastic error-correction models are implemented by means of probabilistic finite-state machines (PFSMs).

Experiments were carried to test our IMT proposal. The results showed that our IMT system outperforms the results of other state-of-the-start IMT systems that are based on word graphs.

Finally, it is worth pointing out that the proposed alternative formalisation of the IMT process can be generalised to obtain a powerful statistical framework that can be applied in other pattern recognition applications such as multi-source translation, computer-assisted speech transcription or computer-assisted transcription of text images.

3.3 Online Learning for IMT

A statistical IMT system able to learn from user feedback by means of the application of online learning techniques is presented. These techniques allow the MT system to update the parameters of the underlying models in real time. Specifically, a log-linear SMT model where all its score components are incrementally updateable is defined. One of the key aspects of the proposed technique is the use of the incremental version of the EM algorithm [9].

The online learning techniques proposed here allow the statistical models involved in the translation process to be updated given the target translations validated by the user. Figure 1 shows a schematic view of these ideas. Here, $\mathbf{s}$ is the source sentence and $\mathbf{t}$ is the target sentence derived by the IMT system from $\mathbf{s}$. By observing $\mathbf{s}$ and $\mathbf{t}$, the user interacts with the IMT system, validating prefixes and/or pressing keys (k) corresponding to the next correct character, until the desired output $\hat{\mathbf{t}}$ is produced. The input sentence $\mathbf{s}$ and its desired translation $\hat{\mathbf{t}}$ can be used to refine the models used by the system. In general, the model is initially obtained through a classical batch training process from a previously given training sequence of pairs $(\mathbf{s}_1, \mathbf{t}_1)$,...,$(\mathbf{s}_n, \mathbf{t}_n)$ from the task being considered. Now, the models can be extended with the use of valuable user feedback by means of online learning techniques.

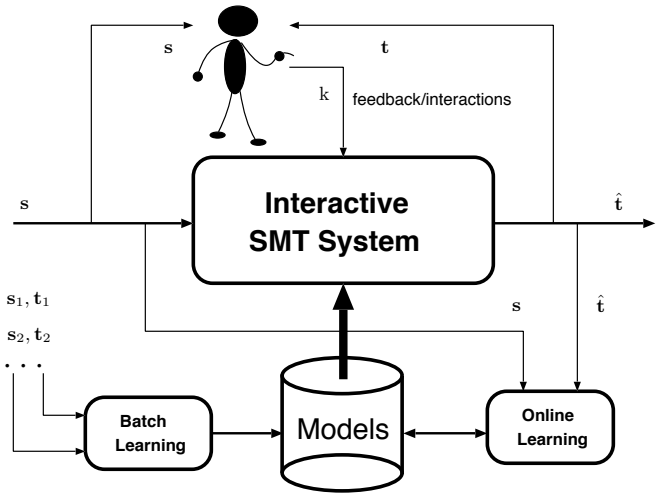


Fig. 1. An Online Interactive SMT system.

According to empirical results, our system significantly outperformed the results of conventional IMT systems. In addition to this, the training times per each new sentence pair allowed the system to be used in a real time scenario.

To the best of our knowledge, the online learning capability that is implemented in our IMT system has never been provided by previous IMT systems. In addition to this, the incremental techniques proposed here can also be exploited to extend SMT systems (in fact, our proposed IMT system is based on an incrementally updateable SMT system). Finally, we think that the use of the incremental version of the EM algorithm has a strong potential utility not only in the field of SMT but also in other fields of natural language processing.

4 Conclusions

This thesis presents several contributions in different aspects of the SMT and IMT fields. Different issues of the three problems to be solved in SMT were addressed, including a new phrase-based model derivation, an alternative estimation technique, a generalised search algorithm that can be parameterized to obtain different algorithms described in the literature, techniques to deal with very large corpora and models both at the estimation and decoding stages, etc. With respect to the field of IMT, two novel techniques able to outperform state-of-the-art sytems have been proposed. In addition to this, we also described an IMT system able to extend its statistical models from user feedback.

5 Publications and Merits

As it was mentioned in section 2.1, this thesis [17] presents both scientific and technologic contributions in the fields of SMT and IMT. As a result of the work on SMT, the open-source THOT toolkit was developed. The toolkit is hosted by SourceForge[1] and released under GPL license. THOT can be downloaded at http://sourceforge.net/projects/thot/. According to the information provided by the SourceForge web site, THOT has been downloaded more than one thousand times since its first release. In addition to this, the toolkit has been cited in several research papers.

Some of the work on SMT as well as the work related to the THOT tolkit was published in an international journal [20] and in an international conference [11]. In addition to this, the author was invited to give a talk on the THOT toolkit at the university of Trento [13].

Additional work on SMT model estimation and search was published in two international conferences [10, 12]. In addition to this, some indirect results using the proposed SMT techniques were published in an international journal [2] and in an international conference [22].

The proposed phrase-level alignment generation technique was published in an international conference [19]. This contribution received the best paper award sponsored by the Springer Verlag editorial to a paper accepted for publication in the EAMT conference.

Work on IMT has been published in different conferences on MT technologies and user interfaces [1, 15, 14, 16, 21], some of them are among the most important ones in their respective fields. Additionally, the IMT system based on error-correction techniques was the basis of two new IMT system proposals which were presented in international conferences [6, 5]. The IMT system with online learning was also the starting point of a new proposal combining IMT with active learning [7]. Finally, the IMT approaches proposed in this thesis were used to build an IMT prototype which is publicly available at http://cat.iti.upv.es/imt/.

In addition to the papers published at conferences and journals, work on IMT developed in this thesis was also published in three book chapters [3, 18, 8].

Finally, the results in this thesis contributed to make progress in two research projects, namely, the I3MEDIA project[2] (PR.068-2007, Spanish CENIT program) and the MIPRCV project[3] (CSD2007-00018, Spanish CONSOLIDER program). In addition to this, the ongoing CASMACAT project[4] (287576, Seventh Framework Programme of the European Union) contains a whole work-package which is focused on online and adaptive learning for IMT.

[1] http://sourceforge.net/
[2] https://i3media.barcelonamedia.org/
[3] http://miprcv.iti.upv.es/
[4] http://www.casmacat.eu/

6 Acknowledgments

Thesis partially supported by the EC (FEDER/FSE), the Spanish Government (MEC, MICINN, MITyC, MAEC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010" CSD2007-00018, iTrans2 TIN2009-14511, erudito.com TSI-020110-2009-439), the Generalitat Valenciana (ALMPR grant Prometeo/2009/014, grant GV/2010/067) and the FPI fellowship.

References

1. Alabau, V., Ortiz, D., Romero, V., Ocampo, J.: A multimodal predictive-interactive application for computer assisted transcription and translation. In: ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces. pp. 227–228. ACM, New York, NY, USA (2009)
2. Andrés-Ferrer, J., Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: On the use of different loss functions in statistical pattern recognition applied to machine translation. Pattern Recognition Letters 29(8), 1072–1181 (2008)
3. Civera, J., González-Rubio, J., Ortiz-Martínez, D.: Multimodal Interactive Pattern Recognition and Applications, chap. Interactive Machine Translation. Springer, 1st edition edn. (2011), http://www.springer.com/computer/hci/book/978-0-85729-478-4
4. García-Varea, I., Ortiz, D., Nevado, F., Gómez, P.A., Casacuberta, F.: Automatic segmentation of bilingual corpora: A comparison of different techniques. In: Proceedings of the Second Iberian Conference on Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, vol. 3523, pp. 614–621. Springer Verlag, Estoril (Portugal) (Jun 2005)
5. González-Rubio, J., Ortiz-Martínez, D., Casacuberta, F.: Balancing user effort and translation error in interactive machine translation via confidence measures. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 173–177 (2010), http://aclweb.org/anthology-new/P/P10/P10-2032.pdf
6. González-Rubio, J., Ortiz-Martínez, D., Casacuberta, F.: On the use of confidence measures within an interactive-predictive machine translation system. In: Proceedings of 14th Annual Conference of the European Association for Machine Translation (2010), http://www.mt-archive.info/EAMT-2010-Gonzalez-Rubio.pdf
7. González-Rubio, J., Ortiz-Martínez, D., Casacuberta, F.: Active learning for interactive machine translation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 245–254 (2012)
8. Leiva, L.A., Alabau, V., Romero, V., Segarra, F.M., Sánchez-Sáez, R., Ortiz-Martínez, D., Rodríguez-Ruiz, L., Serrano, N.: Multimodal Interactive Pattern Recognition and Applications, chap. Prototypes and Demonstrators. Springer, 1st edition edn. (2011), http://www.springer.com/computer/hci/book/978-0-85729-478-4
9. Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Proceedings of the NATO-ASI on Learning in graphical models. pp. 355–368. Norwell, MA, USA (1998)
10. Ortiz, D., García-Varea, I., Casacuberta, F.: An empirical comparison of stack-based decoding algorithms for statistical machine translation. In: New Advance in

Computer Vision. Lecture Notes in Computer Science, Springer-Verlag (2003), 1st Iberian Conference on Pattern Recognition and Image Analysis Mallorca. Spain. June
11. Ortiz, D., García-Varea, I., Casacuberta, F.: Thot: a toolkit to train phrase-based statistical translation models. In: Machine Translation Summit. pp. 141–148. Phuket, Thailand (Sep 2005)
12. Ortiz, D., García-Varea, I., Casacuberta, F.: Generalized stack decoding algorithms for statistical machine translation. In: Proc. on the Workshop on Statistical Machine Translation, HLT-NAACL 2006. pp. 64–71. New York City, USA (Jun 4–9 2006)
13. Ortiz, D., García-Varea, I., Casacuberta, F., Rodríguez, L., Tomás, J.: Thot. New features to deal with larger corpora and long sentences. In: TC-STAR OpenLab on Speech Translation Workshop. Trento, Italy (30th March - 1st April 2006), http://tc-star.itc.it/openlab2006/day1/GarciaVarea.pdf
14. Ortiz, D., Leiva, L.A., Alabau, V., Casacuberta, F.: Interactive machine translation using a web-based architecture. In: Procedings of the International Conference on Intelligent User Interfaces, pp. 423–425. Hong Kong, China (February 2010)
15. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Interactive machine translation based on partial statistical phrase-based alignments. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. Borovets, Bulgaria (sep 2009)
16. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Online learning for interactive statistical machine translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT). pp. 546–554. Los Angeles (Jun 2010)
17. Ortiz-Martínez, D.: Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation. Ph.D. thesis, Universidad Politécnica de Valencia (2011), advisors: Ismael Garca Varea and Francisco Casacuberta
18. Ortiz-Martínez, D., García-Varea, I.: Multimodal Interactive Pattern Recognition and Applications, chap. Incremental and Adaptive Learning for Interactive Machine Translation. Springer, 1st edition edn. (2011), http://www.springer.com/computer/hci/book/978-0-85729-478-4
19. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Phrase-level alignment generation using a smoothed loglinear phrase-based statistical alignment model. In: Proc. of EAMT'08 (2008)
20. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: The scaling problem in the pattern recognition approach to machine translation. Pattern Recognition Letters 29(8), 1145–1153 (2008)
21. Ortiz-Martnez, D., Leiva, L.A., Alabau, V., Garca-Varea, I., Casacuberta, F.: An interactive machine translation system with online learning. In: Proceedings of the ACL-HLT 2011 System Demonstrations. pp. 68–73. Association for Computational Linguistics (2011)
22. Sanchis-Trilles, G., Ortiz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., Hoang, H.: Improving interactive machine translation via mouse actions. In: EMNLP 2008: conference on Empirical Methods in Natural Language Processing. Waikiki, Honolulu, Hawaii (October 25 – 27 2008)

Ph.D Extended Abstract: Session Variability Compensation in Automatic Speaker and Language Recognition

Javier Gonzalez-Dominguez

Biometric Recognition Group (ATVS),
Escuela Politecnica Superior, Universidad Autonoma de Madrid
{javier.gonzalez}@uam.es
http://atvs.ii.uam.es

Abstract. Robust and accurate automatic speaker and language recognition, through the voice signal, remains a challenge for the scientific community mainly due to an old and well-known 'enemy': the session variability, defined as the set of variations among recordings belonging to a same identity (either speaker or language respectively).

During the past decades the issue of compensating/removing undesired variability effects has been broadly accepted as one of the biggest challenges in the field, giving rise to a number of publications full of new manners of somehow avoiding or cleaning the distortions present in the speech signal. However, major advances in the field have not been achieved until the development of new schemes based on Factor Analysis (FA) modelling. This fact responds to the conjunction of several ideas, properly combined in FA, which can be roughly summed up in two key points. First, exploiting prior knowledge in order to model session variability rather than directly removing it; and second, considering session variability as a continuous source rather than a discrete one.

This Ph.D. Thesis is focused on the study, analysis and development of new forms to palliate in a proper way the effects of the session variability problem through recent compensation schemes based on classical FA. In this sense, an extent analysis of the use and mathematical background of FA-based techniques, from the eigen-channels approach to more sophisticated schemes such as Joint Factor Analysis has been conducted.

Further, a special focus has been placed on the use of FA techniques applied to challenging scenarios, as those where the available background data is far from target conditions or the amount of train/test speech is very limited. This is a common case in the increasingly relevant forensic speaker recognition area. Regarding the experimental framework, well-defined and challenging recent automatic speaker and language recognition evaluations (SRE'08 and LRE'09 respectively) have been employed to assess the proposed and studied methods.

Keywords: Speaker Verification (SV), Spoken Language Recognition (SLR), Factor Analysis(FA)

Javier Gonzalez-Dominguez

1 Motivation of the Thesis

Understanding automatic speaker and language recognition systems as valuable tools for different industry and scientific applications, which embrace critical fields as security or forensic apart from others useful applications above mentioned such as data-indexing; and after identifying session variability as the main cause of the performance degradation of this systems, the main motivation of this Thesis [1] is clear: improving automatic speaker and language recognition systems by dealing with session variability. But, more precisely, three observations from the state-of-the-art have mainly motivated the work conducted in this Dissertation. Those being:

- Due to its great ability of dealing with the problem of session variability, a high proliferation of Factor Analysis based methods applied to SV and SLR systems has taken place in a short period of time. However, whereas the basic concepts and hypothesis of the Factor Analysis are widely extended, a small amount of work has been published to deep review the mathematical foundations of Factor Analysis modelling, as well as to clarify the necessary modifications to its integration into SV and SLR fields. This fact has often conducted to a certain obscurity about the implementation process and also to the use of FA tools in a *black-box* mode, without a deep understanding of them.
- Related to the above observation, despite some valuable efforts such as those conducted in specific workshop in the field as JHU 2008[2] and Bosaris 2010[3], little research has been published to put on the same context the different manners to incorporate Factor Analysis into speaker and language verification systems. Even although same protocols or databases are often used by the scientific community, some other systems differences in the configuration parameters among published works (different number of Gaussians, type of features) hinder a fair comparison between the different forms of Factor Analysis.
- The increasing interest in forensic speaker recognition and the need of finding appropriate solutions to the often very adverse session variability conditions associated to this field. Due to the confluence of more robust and accurate systems as well as a better understanding in the field [1] [2], the interest to integrate automatic speaker recognition in the forensic field, to adequately supplement the labour carried out by the expert (eg. phoneticians) has rapidly expanded in recent years. In that sense, a little amount of research [3] [4] [5] has been conducted to explore, analyse and deal with the multiple problems and specificities encountered in the area.

[1] Complete Thesis can be found online in http://atvs.ii.uam.es/files/2011_thesisJavierGonzalez.pdf
[2] http://www.clsp.jhu.edu/workshops/ws08/groups/rsrovc/
[3] http://speech.fit.vutbr.cz/en/workshops/bosaris-2010

2 The Thesis

The Thesis developed in this Dissertation can be stated as follows:

> *Exploiting prior knowledge about speech signal variability, conceived this as a continuous source, to properly include and adapt it to the particular characteristics of the target scenarios is essential to build robust and reliable automatic speaker and language recognition technology.*

3 Objectives

This Dissertation pursuits the following two prime objectives to a major benefit of the automatic SV and SLR systems:

- Provide insight about Factor Analysis as a powerful and efficient tool to deal with the session variability problem in automatic speaker and language recognition.
- Explore and propose different ways to incorporate FA into speaker and language recognition systems, able to obtain significant gains even in very adverse scenarios conditions.

4 Relevant Theory: The Joint Factor Analysis Model

Given a classical GMM system with C Gaussian components and F feature dimensions, where a UBM has been previously trained, it can be seen that, by classical MAP, a speaker-dependent means model supervector $\boldsymbol{\mu}_s$ $(CF \times 1)$ of a new speaker s is derived from the UBM means supervector $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{z}_s \tag{1}$$

where the term $\boldsymbol{D}\boldsymbol{z}_s$ represents the *shift/offset* from the mean $\boldsymbol{\mu}$ as a result of the MAP adaptation, and it is formed by the diagonal $CF \times CF$ matrix $\boldsymbol{D}$, and the $CF \times 1$ weights vector $\boldsymbol{z}_s$ which is assumed to be distributed with a standard normal prior.

By the form in equation 1 and assuming the prior of $\boldsymbol{z}$ standard normal distributed, it can be inferred that, in MAP, speaker-dependent means supervectors are considered to be normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{B} = \boldsymbol{D}^2$, $CF \times CF$. An analogous analysis can be performed with EV-MAP, but considering the variance of the distribution to be confined within a subspace of rank R_s within the supervector space, where $R_s << CF$. Note that the implicit assumption formulated in EV-MAP is then that the eigen-analysis of covariance $\boldsymbol{B}$ results on a few non-zero eigenvalues, exactly R_s. In matrix form

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s \tag{2}$$

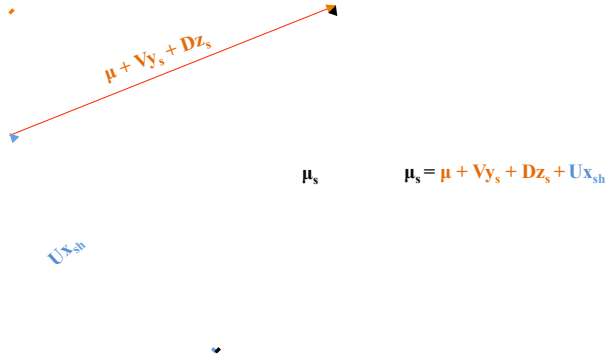


Fig. 1: Representation of the speaker supervector decomposed in the speaker and session variability components.

where $\boldsymbol{V}$ is a low-rank matrix ($CDxR_s$) which explains the speaker variance, in this case $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^T$ and $\boldsymbol{y}_s$ the *weights* which represent the speaker s through the speaker variability subspace spanned by $\boldsymbol{V}$. Note, nevertheless, that by varying $\boldsymbol{y}_s$, the model $\boldsymbol{\mu}_s$ varies across the space spanned by $\boldsymbol{V}$; that is within a R_s-dimensional linear manifold of the supervector space.

JFA integrates both modelling ideas in order to derive the speaker-dependent component of a mean speaker supervector model. So that

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s \tag{3}$$

Note that by this form the assumed variance $\boldsymbol{B}$ is now explained by both $\boldsymbol{V}$ and $\boldsymbol{D}$ ($\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{D}^2$), and as such, it combines the advantages of MAP and EV-MAP: first, the variability is supposed to be, to great extent, constrained in the subspace spanned by $\boldsymbol{V}$; and second, other speaker variability out of this manifold is also accounted. The vector $\boldsymbol{y}_s$ ($R\times 1$) is usually referred to as ***speaker factors***, since represents the speaker variability within $\boldsymbol{V}$, and mathematically responds to the latent factors within a FA modelling as it will be shown later on in the second part of this Chapter.

Once the speaker-dependent component has been established, the session-dependent component of the means speaker supervector is incorporated. By JFA, it is assumed that every utterance h corresponding to a speaker s produces a *distortion* in its speaker mean supervector and this can be modelled via EC-MAP. The supervector space is then modified by and additional term as

$$\boldsymbol{\mu}_{sh} = \boldsymbol{\mu}_s + \boldsymbol{U}\boldsymbol{x}_{sh} \tag{4}$$

where $\boldsymbol{U}$ is a low rank matrix ($CF \times R_c$) that plays the same role than $\boldsymbol{V}$ in EV-MAP but representing the session variability subspace, and $\boldsymbol{x}_{sh}$ is the analogous term of $\boldsymbol{y}_s$. The components of $\boldsymbol{x}_{sh}$ are usually called ***channel factors*** and

Term	*Description*	*Dimensionality*
$\boldsymbol{\mu}$	Mean of the new models. Usually, the UBM speaker mean supervector.	$CF \times 1$
$\boldsymbol{V}$	Speaker Variability Subspace. Low-rank matrix.	$CF \times R_s$
$\boldsymbol{D}$	Besides $\boldsymbol{z}_s$, residual speaker term. Full-rank matrix (diagonal)	$CF \times CF$
$\boldsymbol{U}$	Session Variability Subspace. Low-rank matrix.	$CF \times R_c$
$\boldsymbol{y}_s$	Speaker Factors	$R_s \times 1$
$\boldsymbol{z}_s$	Besides D, residual speaker term	$CF \times 1$
$\boldsymbol{x}_h$	Channel Factors	$R_c \times 1$

Table 1: JFA model components description, (equation 5).

unlike the speaker factors, those depend on the utterance h apart from the speaker s.

Summing up, the Joint Factor Analysis, geometrically represented in Figure 1, is formulated in matrix terms as

$$\boldsymbol{\mu}_{sh} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}z_s + \boldsymbol{U}\boldsymbol{x}_{sh} \tag{5}$$

Table 1 describes each component of the model.

Thus, given a recording or training material h belonging to the speaker s, the JFA model is composed by the tuple of speaker-independent *hyperparameters* $\Lambda = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}\}$, the speaker-dependent *factors* $\boldsymbol{y}_s, \boldsymbol{z}_s$ and the speaker- and utterance-dependent $\boldsymbol{x}_{sh}$ factors. As it will be shown later on, the hyperparameters are pre-trained in a development stage, and remain fixed for all speakers and utterances both in training and testing stages. On the other hand, the set of factors are estimated per each utterance given the speaker-specific data and trained hyperparameters.

At this point, it is convenient to highlight some considerations about the JFA model:

- The JFA model generalizes MAP, EV-MAP and EC-MAP. In fact, the three models can be achieved from JFA by suppressing or zeroing terms (e.g EV-MAP can be obtained by setting to zero $\boldsymbol{U}$ and $\boldsymbol{D}$).
- The factors $\boldsymbol{z}_s$, $\boldsymbol{y}_s$ and $\boldsymbol{x}_{sh}$ are considered to be standard normally distributed $N(0, I)$.
- There is not an analogous term to $\boldsymbol{Dz}$ on the session variability component. Note that if there was, the session component would cover the entire supervector space. This would allow to turn a given speaker into any other just by varying session effects.

5 Experiments on Speaker and Language Recognition

Results presented in this section, in both SV and SLR, summarize some selected results from the complete works contained in the Thesis. Extended results, conclusions and clarifications can be found in the Thesis book. Further, for complete

meaning, those results should be put in the context on an appropriate description of the experimental protocol and databases used (Thesis chapter 5) in the NIST speaker and language recognition evaluation series.

5.1 Speaker Recognition

A comparison among different approaches versus global JFA model is done in this section in the context of speaker verification. Here, the speaker variability is jointly modelled by the classical MAP adaptation and the component provides by the eigenvoice approach $\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{z_s} + \boldsymbol{V}\boldsymbol{y_s}$. As in the other approaches evaluated, the session variability compensation is applied in both training and test recordings in the statistics domain.

	Equal Error Rate (EER in %)			
System (female)	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
Baseline-female	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
MAP-SVC	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
EV-SVC	10.41/0.052	9.24/0.044	9.21/0.045	8.07/0.039
JFA	9.03/0.044	8.22/0.040	8.18/0.041	**7.29/0.039**

Table 2: *Results on SRE'08 tel-tel (female) condition for the four evaluated systems, MAP, MAP-SVC, EV-SVC and JFA.*

	Equal Error Rate (EER in %)			
System (male)	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
Baseline-male	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
MAP-SVC	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
EV-SVC	8.81/0.049	7.33/0.034	7.48/0.037	6.01/0.028
JFA	7.25/0.041	6.26/0.030	6.01/0.031	**5.09/0.025**

Table 3: *Results on SRE'08 tel-tel (male) condition for the four evaluated systems, MAP, MAP-SVC, EV-SVC and JFA.*

Tables 2 and 3 collects the results obtained with the JFA system besides classical MAP adaptation (baseline), MAP adaptation with session varibility compensation (MAP-SVC) and the eigenvoice approach with session variability compensation (EV-MAP), for male and female conditions. Expectedly, the combination of elements outperforms the best results achieved so far for the other systems. The fact of considering the prior represented by the speaker variability subspace but also allowing the speaker variability to lie outside of it, gets to join the advantages provided for both approaches. Improvements of 4% and 4.5% respect to the MAP-SVC system; and a 40%, 48% respect to the MAP systems, are achieved for the female and male conditions respectively.

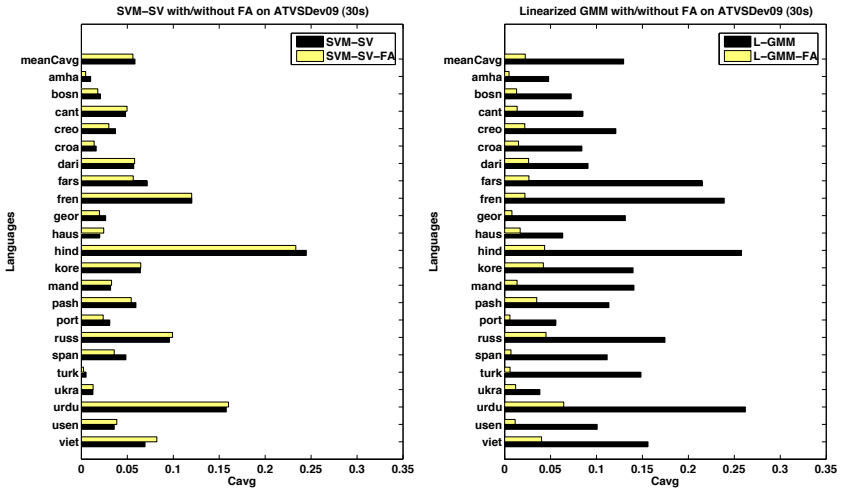


Fig. 2: Effect of session variability compensation on SVM-SV and FA-GMM-LS systems. Results on ATVS-Dev09 using VOA models and U_{VOA}.

5.2 Language Recognition

The need of proper session variability compensation is showed in Figure 2 where both spectral systems, FA-GMM-LS and SVM-SV are assessed with and without compensation via factor analysis on ATVSDev09. Results shows that channel compensation via FA is crucial in GMM modelling performance, getting an improvement of about 82% in $meanC_{avg}$ terms. Also, system SVM-SV take advantage of this compensation but to a lesser extent (4%). This effect appears due to differences in SVM and GMM modelling. In GMM, target languages models, trained with huge amount of data, are far shifted with respect UBM reference model after even a single MAP adaptation. This mean shifting includes not only information belonging to the language but session variability found in the training database which it is mainly independent of the languages. This leads to models that are growing strongly affected by session variability effects. On the contrary, the SVM exhibits a higher robustness to this problem due to its ability to estimate an hyperplane separating target single utterances models against all non-target ones. However, once session variability compensation is applied, both GMM and SVM-SV system, as well as the fusion of both clearly outperforms the performance achieved without session variability compensation via FA.

6 Conclusions

The main conclusions that can be extracted and have been highlighted through this Thesis are:

- The session variability problem is one of the main causes of system performance degradation in both automatic speaker and language recognition systems.
- The session variability should be treated as continuous rather than in a discrete way, since it is the result of the conjunction of a numberless of sources which cannot be properly quantified.
- Most of the session and speaker/language variability associated to a given recording can be explained by a reduced number of variability directions and corresponding weights. Those variability subspaces can be previously estimated from large amount of data and be used as strong priors in the modelling of speaker/language or session variability. This process fits with the theory of Latent Variable models, specifically with Factor Analysis modelling.
- The use of a complex mathematical framework as Factor Analysis is not incompatible with the development of efficient systems. FA can be incorporated in an properly manner in speaker and language recognition, leading to robust and very efficient systems.
- Factor Analysis should not be used as either a closed formula or as a *black box* to deal with session variability. A deep understanding of this modelling strategy as well as the target data (data in operational conditions) nature is needed in order to achieve significant results. A non-adequate use of FA could lead the global system to fail.
- The database mismatch and the short durations problem still being a challenge for speaker and language recognition systems, and although FA can be useful to deal with them, further research is needed to adequate its use in those scenarios.

Main contributions and results are:

- The compilation of the mathematical grounds of Factor Analysis, from its original formulation to its use in speaker and language recognition systems.
- The efforts made in achieving robust but also efficient Factor Analysis based acoustic systems for both speaker and language recognition
- The novel methods explored and proposed to incorporate Factor Analysis into speaker and language recognition systems.
- The study of the main problems in the deployment of speaker and language recognition systems in "real-world" scenarios and the novel methods proposed to mitigate their negative impact in performance by using Factor Analysis.

7 Future Work

A number of research lines arise from the work conducted in this Thesis. Among then, following ones are highlighted:

- Exploring new forms of Factor Analysis applied to palliate the session variability problem and modelling speaker variability. Although JFA has demonstrated to be very effective the, new improved versions could achieve better performance results. Recent strategies as Total Variability [6] o Probabilistic Linear Discriminant Analysis [7] are an example of those evolved FA methods.
- Including a full-Bayesian treatment on Factor Analysis methods. Although, it has been noted that speaker and channel factors (latent variables) has been well modelled rather than make use of a point estimate, other model parameters such as the variability subspaces are estimated via a Maximum Likelihood procedure. This fact could lead to problems as the over-fitting and may be solved via a full Bayesian treatment of all the model parameters involved in FA [8]. Recent work in that sense is accomplished in the field of speaker recognition in [9].
- Exploring new forms to palliate the *short durations* problem via Factor Analysis. Although the problem of short durations is still being a challenge in the field, a scant amount of research has been conducted in the area [10], specially when durations of the recordings vary from one trial to the next, as usual occurs in tasks as forensic speaker recognition. Recent studies as this conducted in [11] endorse this research line.
- Exploring new forms to palliate the *database mismatch* problem via Factor Analysis. Identified as one of the main challenges when dealing with "real-world" systems [3], the database mismatch problem is an open research line where Factor Analysis has not been completely exploited. Initial works have already been conducted in [4] [12].
- Studying the application of those session variability compensation schemes to other biometric recognition traits such as fingerprint or signature verification. The session variability is not a specific problem of speaker and language recognition but in general affects to any biometric trait, as the results of different aspects of their acquisition processes. For instance, the use of different sensors in the capture of fingerprints or signatures includes different session variations that could be faced via Factor Analysis.
- Considering the application of Factor Analysis in automatic speech recognition systems. In the line of the above point, the speech recognizers are strongly affected by a number of variability sources (note that in this case even the inter-speaker variability is considered as a nuisance source). Pioneer experiments in the area have been already conducted in [13] [14].
- Exploring discriminative approaches based on Factor Analysis. Discriminative approaches as SVM has been proved to be very effective in both speaker and language recognition. The idea of derive a discriminative FA model rather than the generative presented in this Thesis is an open line in the development of Factor Analysis based systems. Initial research in this line endorses this future line [15].
- Combining traditional and automatic speaker/language recognition approaches. It is widely agreed upon the scientific community [1] that combining automatic and classical speaker/language recognition approaches [16] [17] should

lead to a major benefit of the recognition systems. Pioneer studies in this field has showed excellent results [18] [19].

8 Publications

Following list enumerates main publications derived from this thesis.

8.1 Book Chapters

1. D. T. Toledano, D. Ramos, **J. Gonzalez-Dominguez** and J Gonzalez-Rodriguez, "Speech Analysis", Stan Z. Li (Eds.), Encyclopedia of Biometrics (ISBN 978-0-387-73003-5), Springer, July 2009.
2. D. Ramos, **J. Gonzalez-Dominguez**, D. T. Toledano and J. Gonzalez-Rodriguez, "Speaker Feature", Stan Z. Li (Eds.), Encyclopedia of Biometrics (ISBN 978-0-387-73003-5), Springer, July 2009.

8.2 Journals indexed (JCR)

1. I. Lopez-Moreno, D. Ramos, **J. Gonzalez-Dominguez** and J. Gonzalez-Rodriguez. Von MisesFisher Models in the Total Variability Subspace for Language Recognition, IEEE Signal Process Letters, 2011, volume 18, number 1, pp 705-708. JCR: 1.165
2. **J. Gonzalez-Dominguez**, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano and J. Gonzalez-Rodriguez. "Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009", IEEE Journal on Selected Topics in Signal Processing, IEEE, 2010, Volume 4, number 6, pp 1084-1093. JCR: 2.647
3. J. Gonzalez-Rodriguez, D. Ramos-Castro, D. Torre-Toledano, A. Montero-Asenjo, **J. Gonzalez-Dominguez**, I. Lopez-Moreno, J. Fierrez-Aguilar, D. Garcia-Romero and J. Ortega-Garcia, "On the use of high-level information for speaker recognition: the ATVS-UAM system at NIST SRE 2005", IEEE Aerospace and Electronic Systems Magazine, Vol. 22, n. 1, pp. 15 - 21, January 2007. JCR: 0.331

8.3 International Conferences

1. J. Gonzalez-Rodriguez, **J. Gonzalez-Dominguez**, J. Franco-Pedroso, D. Ramos. A linguistically-motivated speaker recognition front-end through session variability compensated cepstral trajectories in phone units, in Proceedings of ICASSP 2012, Kyoto, Japan, March 2012.
2. D. Ramos, **J. Gonzalez-Dominguez**, E. Arevalo and J. Gonzalez-Rodriguez. "High-performance session variability compensation in forensic automatic speaker recognition", in Proceedings of 2nd Pan American Meeting on Acoustics. Acoustical Society of America. ISSN 0001-4966., Acoustical Society of America., pp. 2378, November 2010.

3. S. Perez-Gomez, D. Ramos, **J. Gonzalez-Dominguez** and J. Gonzalez-Rodriguez. "Score-level compensation of extreme speech duration variability in speaker verification", in Proceedings of Interspeech 2010, Makuhari, Japan, September 2010.
4. **J. Gonzalez-Dominguez**, B. Baker, R. Vogt, J. Gonzalez-Rodriguez and S. Sridharan. "On the Use of Factor Analysis with Restricted Target Data in Speaker Verication", in Odyssey-10: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 2010.
5. **J. Gonzalez-Dominguez**, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T.Toledano and J. Gonzalez-Rodriguez, "Multilevel and channel compensated language recognition: ATVS system at NIST LRE 2009", in I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language, Porto Salvo, Portugal, September 2009.
6. D. Ramos, J. Gonzalez-Rodriguez, **J. Gonzalez-Dominguez** and J. J. Lucena-Molina. "Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-case database in Spanish", in Proceedings of Interspeech 2008, pp. 1493-1496, September 2008.
7. D. Torre Toledano, **J. Gonzalez-Dominguez**, A. Abejn-Gonzalez, D. Spada, I. Mateos-Garcia, J. Gonzalez-Rodriguez. Improved language recognition using better phonetic decoders and fusion with MFCC and SDC features. INTERSPEECH 2007: 194-197

8.4 National Conferences

1. **J. Gonzalez-Dominguez**, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano and J. Gonzalez-Rodriguez, "ATVS-UAM NIST SRE 2010 System", in Proceedings of FALA 2010, Vigo, Spain, November 2010.
2. **J. Gonzalez-Dominguez**, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano and J. Gonzalez-Rodriguez, "Eficiencia Computacional y Alto Rendimiento en Reconocimiento Automtico de Locutor: el Sistema ATVS-UAM en NIST SRE 2010 ", in Actas de las V Jornadas de Reconocimiento Biomtrico de Personas, Huesca, Spain, September 2010.
3. Montero-Asenjo, **J. Gonzalez-Dominguez**, D. Ramos-Castro, I. Lopez-Moreno, D. T. Toledano and J. Gonzalez-Rodriguez. "On the use of high-level information in speaker and language recognition", in Proceedings of the Spanish Network of Speech Technologies Workshop (ISBN 84-96214-82-6), pp. 15-18, Zaragoza, Espaa, November 2006.

9 Awards

The author was awarded with a PhD scholarship from Ministerio de Educacion y Ciencia between 2007 and 2011, which partially supported the research summarized in this Dissertation.

The author was awarded with the Microsoft Research Best Student Paper Award at international conference SIG-IL/Microsoft 2009, for one publication

originated from this Dissertation: "Multilevel and Channel-compensated Language Recognition: ATVS-UAM systems at NIST LRE 2009", Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano e Joaquin Gonzalez-Rodriguez, Proc. of SIG-IL/Microsoft 2009, Porto Salvo, Portugal, September 2009.

Derived from partial work of this dissertation, the author was granted with an intern scholarship from Google Inc. This internship was entirely developed at the speech recognition technologies department of Google Research, NY, U.S.A., under the supervision of Pedro Moreno and Eugene Weinstein during the period July-October 2010.

Derived from partial work of this dissertation, the author has been invited to give talks in the area in prestigious speech labs as SRI international, Menlo Park, U.S.A and ICSI, Berkeley, U.S.A.

Acknowledgments. This research has been supported by the Ministerio de Ciencia e Innovacion under the proyect TEC2009-14719-C02-01 and Catedra UAM-Telefonica.

References

1. J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
2. D. Ramos, "Forensic evaluation of the evidence using automatic speaker recognition systems," Ph.D. dissertation, Universidad Autonoma de Madrid, November 2007.
3. D. Ramos, J. Gonzalez-Rodriguez, and J. Gonzalez-Dominguez, J. Lucena, "Adressing Database Mismatch in Forensic Speaker Recognition with Ahumada III: A Public Real-Casework Database in Spanish," in *Interspeech*, 2008.
4. J. Gonzalez-Dominguez, B. Baker, R. Vogt, J. Gonzalez-Rodriguez, and S. Sridharan, "On the Use of Factor Analysis with Restricted Target Data in Speaker Verification," in *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010.
5. D. Ramos, J. Gonzalez-Dominguez, E. Arevalo, and J. Gonzalez-Rodriguez, "High-Performance Session Variability Compensation in Forensic Automatic Speaker Recognition," in *Proceedings of 2nd Pan American Meeting on Acoustics. Acoustical Society of America. ISSN 0001-4966.* Acoustical Society of America., November 2010, p. 2378.
6. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788 – 798, February 2011.
7. P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010.
8. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, Oct. 2007. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0387310738

9. J. Villalba and N. Brummer, "Towards Fully Bayesian Speaker Recognition: Integrating out the between-Speaker Covariance," in *INTERSPEECH*, 2011.
10. S. Perez-Gomez, D. Ramos, J. Gonzalez-Dominguez, and J. Gonzalez-Rodriguez, "Score-Level Compensation of Extreme Speech Duration Variability in Speaker Verification," in *Proceedings of Interspeech 2010*, September 2010.
11. M. I. Mandasari, M. McLaren, and D. van Leeuwen, "Evaluation of i-vector Speaker Recognition Systems for Forensic Application," in *INTERSPEECH*, 2011.
12. M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An I-Vector Extractor Suitable for Speaker Recognition with Both Microphone and Telephone Speech," in *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010.
13. D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian Mixture Models for Speech Recognition." in *ICASSP*, 2010, pp. 4330–4333.
14. L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, "Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models." in *ICASSP 10*, 2010, pp. 4334–4337.
15. O. Glembek, L. Burget, N. Brummer, O. Plchot, and P. Matejka, "Discriminatively trained i-vector extractor for speaker verification," in *INTERSPEECH*, 2011.
16. H. J. Kunzel, "Current Approaches to Forensic Speaker Recognition," in *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 1994, pp. 135–141.
17. P. Rose, "Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 159–191, 2006.
18. A. de Castro, D. Ramos, and J. Gonzalez-Rodriguez, "Forensic Speaker Recognition Using Traditional Features Comparing Automatic and Human-in-the-Loop Formant Tracking," in *INTERSPEECH*, 2009, pp. 2343–2346.
19. J. Gonzalez-Rodriguez, "Speaker Recognition Using Temporal Contours in Linguistic Units: The Case of Formant and Formant-Bandwidth Trajectories," in *INTERSPEECH*, 2011.

Detection and Handling of Overlapping Speech for Speaker Diarization

Martin Zelenák and Javier Hernando

TALP Research Center,
Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya
C. Jordi Girona 31, 08034 Barcelona, Spain
martin.zelenak@upc.edu

Abstract. This thesis concerns the detection of overlapping speech segments and its further application for the improvement of speaker diarization performance. We propose the use of three spatial cross-correlation-based parameters for overlap detection on distant microphone channel data. Spatial features from different microphone pairs are fused by means of principal component analysis or by an approach involving a multilayer perceptron. In addition, we investigate the possibility of employing long-term prosodic information. The most suitable subset of candidate prosodic features is determined by a two-step mRMR feature selection algorithm. For segments including detected overlapping speech the speaker diarization system picks a second speaker label, and such segments are also discarded from the model training. The proposed overlap labeling technique is integrated in the Viterbi-decoding part of the diarization algorithm.

Keywords: overlapping speech detection, speaker overlap, speaker diarization, spatial features, cross-correlation, prosody

1 Introduction

It is a well known fact that people sometimes tend to speak at the same time, i. e., simultaneously. It is a normal part of human conversation behavior. For that reason, audio recordings of meetings commonly include regions of overlapping speech. A lot of spoken language technologies suffer from this conversation phenomenon, one of them is speaker diarization.

Given a speech recording, *speaker diarization* aims to answer the question: *"Who spoke when?"*, generally without any prior knowledge. The application of diarization systems is often a very useful preprocessing step for other audio technologies such as ASR. Meetings are considered the most difficult application domain due to high spontaneity of speech, variable microphone signal quality, and room reverberation.

According to several studies [1, 2], a portion of the performance degradation on real meeting data can be directly associated with the occurrence of speaker

overlaps. Nevertheless, the number of published proposals is rather limited and dealing with overlapping speech still remains a challenging problem.

The discussed thesis[1] addresses the issues related to the occurrence of simultaneous speech in meeting recordings. The motivation is to improve speaker diarization performance. However, the investigation of overlapping speech may also be useful beyond diarization, e.g., for speech, or speaker recognition. This paper briefly summarizes the outcomes of the thesis work and highlights some of the achieved results.

2 Motivation and Thesis Objectives

A common drawback of conventional diarization systems is that they are only able to assign one speaker label per segment. In cases when a segment contains simultaneous speech, this implicitly leads to missed speech errors. In addition, another possible effect of overlaps on diarization performance discussed in [1] is that speaker models could be corrupted if simultaneous speech is included into the training data.

The first main goal of the thesis was the development of a robust overlap detection system. The requirement was to work with distant channel data without any constraints about microphone configuration or the recording room. Our interest was to research and propose new features which may be useful for this task. We aimed at exploring the possibilities of employing spatial-based information for the detection of simultaneous speech since (smart) meeting rooms are normally equipped with microphone arrays. The availability of multi-channel data provides the option to estimate features that are in some way related to spatial location. Further option was to investigate the potential of higher-level information. "Higher" in this case refers to speech information which is above the level of short-term spectral or cepstral features, such as prosody.

Another main goal of this thesis was to apply the detected overlapping speech in the UPC speaker diarization system in order to reduce diarization error. This should be achieved by both recovering missed speaker time, as well as by improving the clustering. We sought to implement a novel technique for the assignment of extra speaker labels in speaker overlap segments. Different overlap detection systems were examined according to the quality of their hypotheses for diarization improvement.

Finally, since our general intention was contribute to the research in human language processing, we participated in the organization of the Albayzin evaluation campaign. Being in charge of the speaker diarization section, we presented the final results in [3] and [4].

3 Overlapping Speech Detection

Our baseline overlap detection system was firstly defined in [5], it utilizes a number of spectral-based features, such as MFCCs, LPC residual energy [6], spectral

[1] Thesis manuscript is accessible at: `http://www.tdx.cat/handle/10803/72431`

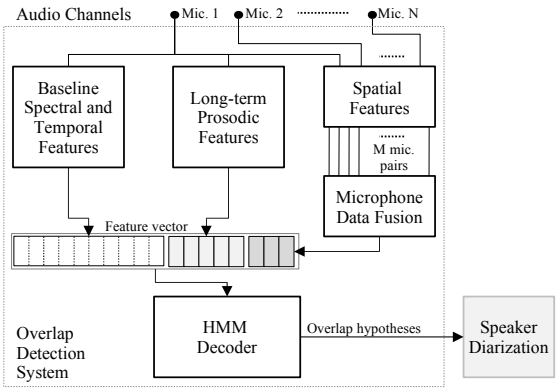


Fig. 1. Overlap detection system diagram.

flatness [7], and deltas. Following classes are considered to classify the signal and produce an output hypothesis by means of Viterbi decoding: non-speech (e. g., silence, noise), single-speaker speech, and overlapping speech. For each class a three-state hidden Markov model (HMM) is defined. Since the amount of training data is not balanced among classes, we use a higher number of Gaussian components for single-speaker speech than for overlapping speech and non-speech. Transition probabilities between different HMMs are set manually. In order to increase the precision, the transition from single-speaker speech to overlapping speech can be penalized with an overlap insertion penalty (OIP), and the transitions from overlapping speech to non-speech and vice versa are forbidden. The diagram showing the system architecture is given in Fig. 1.

3.1 Novel Spatial-based Features

The generalized cross-correlation function with phase transform weighting (GCC-PHAT) is a commonly used measure of the similarity between signals performing well in reverberant environments [8]. GCC-PHAT exhibits a prominent peak at the elapsed time corresponding to the dominant sound source in the room, minimizing the peaks of the non-dominant sources and reverberation at the same time.

Our hypothesis was that in case of multiple, possibly moving, concurrent speakers, the time delay of arrival (TDOA) estimates produced by the GCC-PHAT will jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice. TDOA can be expressed as follows,

$$\hat{\tau}_{mn} = \arg\max_{\tau} R_{mn}(\tau), \tag{1}$$

where $R_{mn}(\tau)$ is the GCC-PHAT function for a pair of microphones m and n. The maximum value of the cross-correlation sequence should also be lower than in the single speaker situation, since multiple speakers introduce random

peaks, which attenuate the main peak. We proposed three cross-correlation-based spatial features for every microphone pair, which are intended to provide some degree of information on speaker overlaps [5, 9].

The first is the *coherence value*, defined in (2), which is the principal peak value of the GCC-PHAT. In ideal conditions it should be high for single-source situations, while the presence of noise, reverberation, and concurrent acoustic sources attenuate this value.

$$C_{mn} = \max(R_{mn}(\tau)) \tag{2}$$

The second feature, coherence *dispersion ratio*, is derived from the coherence value. It is defined as follows,

$$D_{mn} = \frac{C_{mn}^2}{\sum_{t=-w_{mn}}^{w_{mn}} R_{mn}^2(t + \hat{\tau}_{mn})}. \tag{3}$$

This value is computed as the ratio between the square of the main peak value and the square quadratic sum of the cross-correlation values under a time delay window w_{mn}. The size of the window w_{mn} varies for different microphone pairs and it is set to the TDOA standard deviation of each pair. In this way, the dispersion ratio measures the relation between the energy of the main peak and the energy that is scattered in its neighborhood. Similar to the coherence feature (2), the dispersion ratio is close to 1 in the case of a single speaker and ideal conditions, while it has a lower value in reverberant conditions or concurrent acoustic sources situations.

Finally, the *delta of TDOA* obtained by (1) for every microphone pair also carries information on overlaps. The derivative of the TDOA is high in situations where the speaker is moving, multiple non-concurrent speakers change turns at talk, or multiple speakers talk simultaneously.

3.2 Microphone Data Fusion

Practical issues with the use of spatial features are the high and variable dimensionality of vectors. Especially the latter—sites may have different number of microphones—makes it difficult to train a general model. One of the strategies for dimensionality reduction and normalization is the application of a principal component analysis (PCA), which transforms the original feature space into a new coordinate system with the greatest variance lying on the first component. We estimated a separate transformation matrix for every discussed spatial feature kind per each site, and then we use just the first principal component.

We also considered an alternative approach to reduce the spatial vector dimensionality based on a multi-layer perceptron (MLP). The input of the MLP is composed by 6 input neurons, 3 for spatial features and 3 for normalization values (mean of coherence, variance of coherence, variance of TDOA) for every pair. The output is a binary score classifying between overlap and non-overlap, which is commensurable across microphone pairs. For a given frame the average score was taken and merged with corresponding spectral feature vectors.

3.3 Prosodic Features

Prosody, in general, is characterized by rhythm, intonation, stress, and juncture in speech. A few studies were published that researched the relationship between prosodic cues and the interaction of conversation participants. For instance, it was suggested that stretches of low pitch can trigger backchannel feedback from listener (*yeah, uh-huh, right*) [10]. It was also shown that speakers raise their voices when starting their utterance during somebody else's talk, compared to starting in silence [11].

For the detection of overlapping speech a number of prosody-based features are considered [12, 13]. They can be assigned to one of the following categories: pitch, intensity and (four) formant frequencies. In addition, for each of these categories we estimate besides the actual value for every given time point also long-term statistics such as mean, median, min., max., std. deviation and the difference between the min. and max. value. Missing values, such as F0 estimates for unvoiced speech, or parameters in non-speech regions, are substituted with default values. Non-speech regions are not considered for the computation of the statistical parameters.

The use of both instantaneous prosodic features and their long-term statistics is a source of redundancy in our prosodic feature set. Due to this fact, we have added a two-stage feature selection procedure to our feature extractor. The stages are as follows. In the first stage, we applied a mRMR algorithm [14] on held-out development data to score individually the candidate features against the target class (overlapping vs. single-speaker speech) and sorted them according to their minimum redundancy and maximal relevance.

The second feature selection stage involves conventional hill climbing wrapper approach, i. e., iteratively adding candidate features to the feature set, creating a model and evaluating the system on the development data.

3.4 Overlap Detection Experiments

Experiments were conducted on the AMI corpus, on far-field microphone array channels. We defined a single- and a multi-site scenario. The first included recordings only from Idiap site and the latter also from Edinburgh and TNO site. The average amount of overlap in these scenarios was 14.40% and 15.10%, respectively. Training of the overlap detection system and evaluation is performed with force-aligned annotations obtained by SRI's DECIPHER recognizer.

Performance is measured with recall (ratio between true detected and reference overlap time), precision (ratio between true and all detected overlap time), and with the sum of missed and false overlap time divided by the reference overlap time, referred to as detection error. To make a reasonably fair evaluation, results are measured at four operation points defined by the OIP value ($OIP = \{0, -10, -50, -100\}$).

The detection performance for different feature combination setups is given in Fig. 2. For single-site data it can be seen that the overlap detection also using spatial or prosodic parameters outperformed the baseline system (*Spct*).

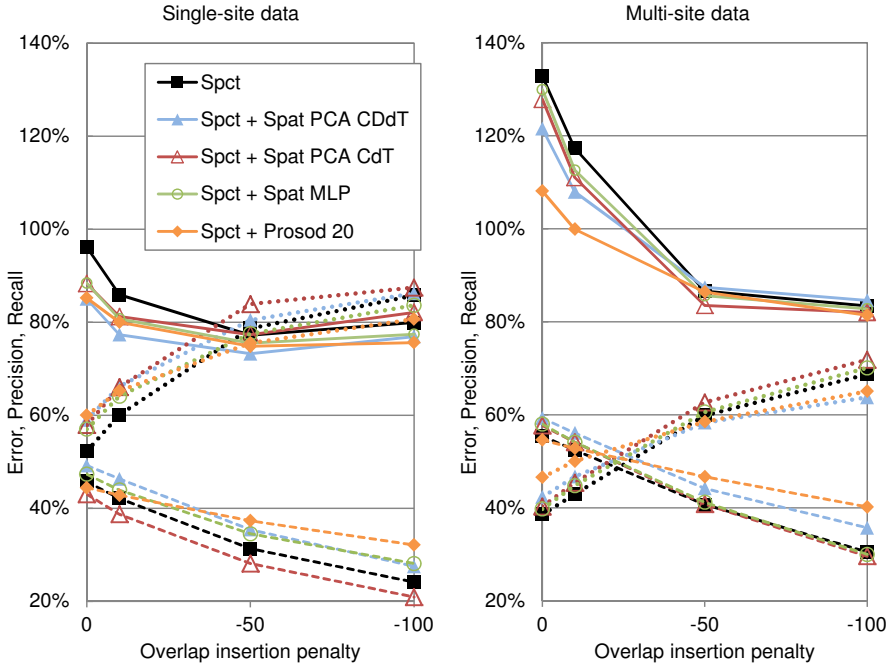


Fig. 2. Overlap detection performance for AMI (left) single- and (right) multi-site data using spectral features (Spct), and in combination with spatial (+ Spat) or prosodic (+ Prosod 20) features. Detection error, precision, and recall are delineated with solid, dotted, and dashed line, respectively.

We can conclude that the PCA-fused spatial feature including dispersion ratio (*Spct + Spat PCA CDdT*) is well suited for the single-site condition. The system using MLP score (*Spct + Spat MLP*) has a good detection performance in this scenario, but for high OIP its precision drops bellow the one of the baseline system.

For the multi-site scenario, however, the mentioned PCA-fused dispersion ratio (marked with *D*) seems to lack robustness. The possible reason for the worse performance of feature setups involving this parameter is its dependency on the spatial distribution of microphones, which might be an issue in case of using multiple recording rooms. Moreover, the limited ability to compensate for the variability of this scenario can most likely be attributed to the simplicity of the PCA technique.

In the multi-site case the better performing setup included only spatial coherence and delta TDOA (*Spct + Spat PCA CdT*), but the distinction in performance between setups including and not including dispersion ratio becomes evident only at higher OIPs. The MLP technique combines all three spatial parameters more effectively in this scenario and outperforms, or at least equals, the baseline system. In general, the less precise multi-site models need a higher

amount of overlap penalization to arrive to the lowest detection errors. The addition of prosodic features decreased the overlap detection error in both scenarios either due to higher precision for low penalties or due to improved recall in high penalty region.

4 Speaker Diarization

Our speaker diarization system, detailed in [15], follows the commonly used agglomerative clustering approach and relies on 20 MFCCs extracted from 30 ms frames. The algorithm starts with an uniform initial segmentation where the number of initial clusters is determined automatically. Clusters are modeled with Gaussian mixture models (GMMs) and cluster pair merging in each iteration is driven by Bayesian information criterion (BIC). The complexity of GMMs is also determined automatically based on the amount of data corresponding to a particular cluster.

The system can be improved by multi-channel approach based on conventional techniques. We applied speech signal techniques such as Wiener filtering and beamforming for signal enhancement, and we also combined the time-delay-of-arrival (TDOA) information as a second stream in the diarization [16].

The performance of the speaker diarization was evaluated by means of the diarization error rate (DER). Defined by NIST, the DER is a time-weighted metric composed of the sum of missed speaker time, false alarms and speaker error time.

4.1 Handling Overlapping Speech

Overlap handling in diarization comprises the labeling and/or exclusion of simultaneous speech. The first technique seeks to select the two most likely clusters in Viterbi decoding instead of only one. In this way the missed speaker time should be decreased. Overlap exclusion blocks overlap frames from being included into cluster initialization and HMM training, but does not prevent decoding them. The aim of this technique is to get lower speaker detection error rates with more precise clusters. The concept is depicted in Fig. 3.

Both techniques work independently from each other, or better said sequentially, since the exclusion works throughout the diarization process whereas the labeling is performed at the end of the iteration process. These two techniques do not necessarily have to share the same overlap hypothesis, they can be optimized independently and can possibly use two different hypotheses, i. e., one for each technique. This method was firstly suggested in [5].

4.2 Speaker Diarization Experiments

Baseline DERs and relative improvements by applying overlap exclusion and labeling in experiments on AMI data are given in Table 1. The most successful overlap detection setup on single-site data was the combination of spectral

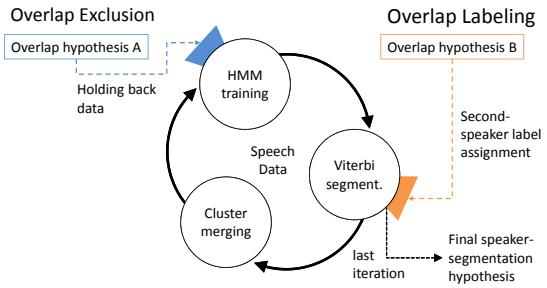


Fig. 3. Overlap handling concept in speaker diarization algorithm.

and the three PCA-transformed spatial parameters (*Spct+Spat PCA CDdT*), yielding improvement of 11.2% relative. A good result was also obtained with the combination of spectral and prosodic features. However, the result was not much higher compared to the *Spct* case.

In the multi-site scenario the relative improvements were higher, the best observed result of 17.0% DER reduction was with the combination of spatial coherence and delta TDOA (*Spct+Spat PCA CdT*). Another successful system with 13.9% improvement was the one using MLP for the spatial parameter fusion.

Table 1. Speaker diarization with exclusion and labeling of simultaneous speech detected by different systems, baseline DER and relative improvements over the diarization baseline (in %).

Overlap det.	Single-site +Ovlp. Excl.	Multi-site and Labl.
Baseline	38.3	37.3
Spct	+6.9	+6.7
Spct+Spat (PCA) CdT	+5.7	+17.0
Spct+Spat (PCA) CDdT	+11.2	+8.0
Spct+Spat MLP	+5.7	+13.9
Spct+Prosod 20	+7.2	+11.1

Table 2 shows the comparison of our Viterbi-integrated labeling technique to two simple labeling schemes in terms of relative DER improvement. The first of these techniques a posteriori attributes the overlapping speaker label according to the nearest neighboring (NNeigh.) speaker, as in [17]. The other competing technique assigns the overlapping label to the most talkative speaker (MTalk.) [18]. In case the most talkative speaker has already been picked by the diarization system, the second most talkative speaker is selected. In general, the differences between DER of the three labeling techniques are small, but it can be seen that the results of the technique proposed in this thesis are competitive, in single-site scenario in particular.

Table 2. Relative improvements of speaker diarization by different labeling strategies.

	Single-site			Multi-site		
Overlap det.	+Labl. Vit.	NNeigh.	MTalk.	Vit.	NNeigh.	MTalk.
Spct	**+5.3**	+5.1	+4.8	**+2.7**	+3.2	+2.3
Spct+Spat	**+6.2**	+5.9	+5.6	**+3.4**	+3.9	+3.0

5 Discussion

This work deals with the issues of overlapping speech in the context of speaker diarization on distant microphone channels. In order to locate the regions where multiple speakers are speaking simultaneously, an overlap detection system was built. We have found that spatial information can be utilized to perform this detection and proposed three novel cross-correlation-based features. The problem of high and variable dimensionality of spatial feature space was addressed with the application of a per-site-specific PCA, or an MLP neural network. Furthermore, we have also introduced features based on prosody and their long-term statistics.

Objectively, the overlap detection performance has still a lot of potential for improvement. The task proved to be extremely challenging for an automated system, but in a lot of cases it is difficult even for humans to decide what can and what cannot be considered overlapping speech.

By handling of the detected simultaneous speech segments, we managed to improve the baseline speaker diarization system. With the objective to build more precise speaker models, the speech frames including overlapping speech were excluded from the training process. In addition, we reduced diarization's missed speech by assigning second speaker labels for speaker overlap segments. Analyses beyond the scope of this paper showed that the performance of overlap exclusion exhibits a relatively non-stable nature [9]. Our proposed labeling technique delivers competitive results compared to alternative simple strategies, especially in the single-site scenario.

References

1. Otterson, S., Ostendorf, M.: Efficient use of overlap information in speaker diarization. In: Proc. ASRU '07, Kyoto, Japan (2007) 683–686
2. Huijbregts, M., Wooters, C.: The blame game: Performance analysis of speaker diarization system components. In: Proc. Interspeech '07, Antwerp, Belgium (2007)
3. Zelenák, M., Schulz, H., Hernando, J.: Albayzin 2010 Evaluation Campaign: Speaker Diarization. In: Proc. FALA 2010, Vigo, Spain (2010) 301–304
4. Zelenák, M., Schulz, H., Hernando, J.: Speaker Diarization of Broadcast News in Albayzin 2010 Evaluation Campaign. EURASIP Journal on Audio, Speech, and Music Processing **2012**(19) (2012)
5. Zelenák, M., Segura, C., Hernando, J.: Overlap Detection for Speaker Diarization by Fusing Spectral and Spatial Features. In: Proc. Interspeech '10, Makuhari, Japan (2010) 2302–2305

6. Boakye, K., Trueba-Hornero, B., Vinyals, O., Friedland, G.: Overlapped speech detection for improved speaker diarization in multiparty meetings. In: Proc. ICASSP '08, Las Vegas, NV, USA (2008) 4353–4356
7. Yantorno, R.E., Krishnamachari, K.R., Lovekin, J.M., Benincasa, D.S., Wenndt, S.J.: The spectral autocorrelation peak valley ratio (SAPVR) – a usable speech measure employed as a Co-Channel detection system. In: Proc. IEEE International Workshop on Intelligent Signal Processing (WISP), Budapest, Hungary (2001)
8. Brandstein, M., Silverman, H.: A robust method for speech signal time-delay estimation in reverberant rooms. In: Proc. ICASSP '97, Munich, Germany (1997) 375–378
9. Zelenák, M., Segura, C., Luque, J., Hernando, J.: Simultaneous Speech Detection with Spatial Features for Speaker Diarization. IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on New Frontiers in Rich Transcription **20**(2) (Feb. 2012) 436–446
10. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics **32/2000** (2000) 1177–1207
11. Shriberg, E., Stolcke, A., Baron, D.: Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech. In: Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ, USA (2001) 13–16
12. Zelenák, M., Hernando, J.: The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization. In: Proc. Interspeech '11, Florence, Italy (2011) 1041–1044
13. Zelenák, M., Hernando, J.: Speaker Overlap Detection with Prosodic Features for Speaker Diarization. IET Signal Processing (2011) in review.
14. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8) (Aug. 2005) 1226–1238
15. Luque, J., Anguera, X., Temko, A., Hernando, J.: Speaker diarization for conference room: The UPC RT07s evaluation system. In Stiefelhagen, R., Bowers, R., Fiscus, J., eds.: Multimodal Technologies for Perception of Humans. Volume 4625/2008 of LNCS. Springer Berlin / Heidelberg (2008) 543–553
16. Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech, and Language Processing **15**(7) (2007) 2011–2022
17. Huijbregts, M., van Leeuwen, D., de Jong, F.: Speech Overlap Detection in a Two-Pass Speaker Diarization System. In: Proc. Interspeech '09, Brighton, UK (2009) 1063–1066
18. van Leeuwen, D.A., Konečný, M.: Progress in the AMIDA Speaker Diarization System for Meeting Data. Multimodal Technologies for Perception of Humans **4625/2008** (2008) 475–483

Phase in the harmonic models of the speech signal: strategies for representation, processing and applications

Ibon Saratxaga, supervised by Inma Hernaez

Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Bilbao
ibon@aholab.ehu.es, inma@aholab.ehu.es

Abstract. This PhD dissertation was written by Ibon Saratxaga and supervised by Inma Hernáez. It was defended at the University of the Basque Country the 5th of March 2012. The committee members were Prof. José Manuel Pardo (UPM), Prof. Asunción Moreno (UPC), Dr..Eduardo Rodríguez Banga (UVigo), Dr. Eva Navas (UPV/EHU) and Dr. Daniel Erro (UPV/EHU).The dissertation was awarded an "apto cum laude" qualification.

Keywords: Phase, harmonic models, polarity, ASR, speaker recognition, speech perception.

1 Introduction

Phase spectrum information of the speech signal has traditionally been neglected by many applications in favour of the information derived from the spectral module. Phase information poses some difficulties from the signal processing viewpoint, due to its special characteristics like time and frequency dependency, phase-wrapping and representation difficulties. Moreover, phase spectrum seems to have little perceptual importance for the human hearing, its influence being of second order compared to the signal module spectrum.

These reasons and the fact that the spectral module is easy to understand and manipulate and it is directly related to perception, have favoured the exclusive use of the module spectrum both in synthesis applications (speech synthesis, voice conversion and transformation, for instance), and in analysis applications (speech and speaker recognition, among others). Let's review the role of phases in these two areas.

1.1 Phases and speech synthesis

In the speech synthesis area phases are often regarded as a problem instead of as a useful feature of the speech. In the case of sinusoidal [1][2] and harmonic models [3][4], for instance, the instantaneous phases of the sinusoidal components are calculated. Instantaneous phase depends on the time of analysis, so using it often implies further processing to assure that no phase mismatches between frames occur, especially when the signal is manipulated (pitch or time modifications) or concatenated.

To solve these problems, the linear phase term (which depends on the analysis instant) has to be removed from the instantaneous phase. Several techniques have been proposed to do this, e.g.: pitch synchronous analysis, calculation of constant points in every pitch period (onset times [5], gravity centres[6] or glottal closure instants).

Other models for speech synthesis, like those based on LPC, directly disregard the phase information and use artificial phase (usually minimum phase but also random or zero phase) to generate the synthetic speech.

These two approaches have direct consequences in the waveform shape of the resulting signal: changing the original phases leads to losing the original waveform shape. The actual importance of keeping the signal shape invariant is not definitely stated. Several authors [5],[7] have proposed shape invariant speech signal manipulation methods, but the effect of the phase manipulation (which leads to waveform shape modification) in the perceptual quality of the speech is not well studied.

The perceptual importance of the phases has been well studied in areas such as electro acoustics and hearing physiology, in general using artificial or musical sounds. In the area of speech technologies, phase contribution to the intelligibility has been evaluated by several authors [8], [9]. Phases have also been investigated from the vocoder point of view to determine how much of the phase information can be discarded with no or little perceptual effect in the transcoded signal [10], [11]. But very few authors have evaluated the effect of modifying the phases in the overall quality of the audio signal: the experiments usually imply heavy phase modifications (fixed or minimum phase) and synthetic signals instead of real speech (e.g. [12]).

Summing up, the role of the phase in the speech synthesis domain can be described as uncertain: phases are tough to manipulate, are often discarded and it is not clear if they have a relevant paper in the resulting speech quality. This thesis tries to bring some light into these questions.

1.2 Phases and speech analysis

Short time spectral information plays a fundamental role in speech analysis, as it is in the base of most of the speech modelling techniques. However, in speech or speaker recognition applications, usually only spectral magnitude is used, while phase information is directly discarded. Phase lacks apparent structure or patterns and this is one reason not to use it for modelling purposes.

Nonetheless, there have been several attempts to include phase information in recognition applications. Group delay related information has been used both in automatic speech recognition (ASR) [13], [14] and in speaker recognition models [15], [16], [17]. In this last area several authors report significant benefits when phase information is combined with magnitude (MFCC) based parameters [18], [19], [20].

One area in speech analysis that is still an open problem is the detection of synthetic impostors, that is, synthetic speech which imitates a speaker's voice who is authorized in a speaker verification system. Current speaker verification systems have serious problems detecting high quality synthetic voice as synthetic and thus rejecting it as false claimant [21], [22]. Phase information could help to discriminate synthetic

speech in many cases, namely those where the original phases have been disregarded or severely altered.

2 Motivation and Objectives

In the framework of the harmonic models for the speech, we have proposed a new representation for the phase information whose features simplify its processing. The RPS (Relative Phase Shift) transformation converts the instantaneous phase of the spectrum to a measure of the relative phase shift between each harmonic component and the fundamental frequency.

This transformation eliminates the dependency of the phase with the frequency and the analysis instant. The RPSs are only dependent on the initial phase shifts between components, and thus, their interpretation and manipulation is much easier than the instantaneous phase's one. This representation is convenient for phase manipulation in synthesis using harmonic models, but more important, it makes the signal phase structure evident, as can be seen in the following figure.

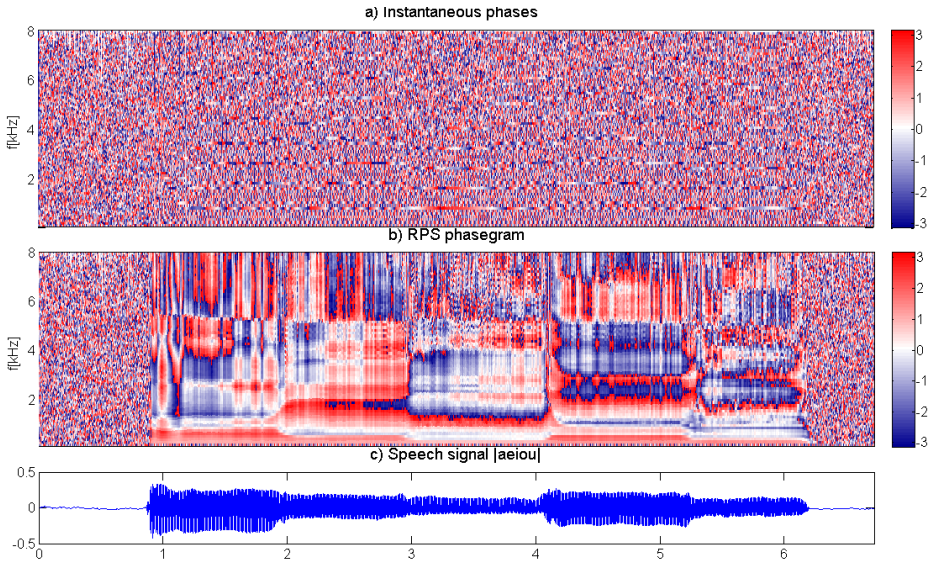


Fig. 1. Instantaneous phase (a) and RPS (b) phasegrams for the signal /aeiou/ (c)

Fig. 1 shows two phasegrams (similar to the spectrograms, but showing phase instead of module) of a speech signal with sustained vowels /aeiou/. The difference between (a) where the instantaneous phases do not show any structure and the clear phase patterns that arise in the RPS phasegram (b) is evident.

The main motivation of the thesis is to study the features and potential applications of this new phase representation, the RPS. Therefore, the main objectives of the thesis can be summarized as follows.

2.1 Objectives

The principal objective of the thesis has been to analyse and understand the potential of the phase information represented by means of the RPS, in different areas of the speech signal processing domain. This general objective was split into more specific ones:

- Phases in the voice signal characterization: Interpretation of the RPS and analysis of the relationship between their features and the voice characteristics like formants, polarity or source and vocal tract separation.
- Phase modelling: Finding a parameterization for the phase information suitable for statistical modelling.
- Phases in the speech signal analysis: Research on the influence of phoneme and speaker features on the phase information of the signal. Application in several tasks in the speech analysis area: speech recognition, speaker recognition and synthetic speech detection.
- Phases and perception: Evaluation of the perceptual influence of the phase manipulations, trying to answer to the question of the real importance of implementing phase handling strategies in speech synthesis.

3 Methodology

Due to the broadness of the objectives of the thesis, the research has been divided in several separate stages with different methodologies employed in each of them. Generally speaking, for each of the areas that have been covered in the thesis we have reviewed the relevant literature, focusing both in the state of the art of the area and in the reported applications of the phase information. Then, depending on the case, we analyse any relevant feature of the RPS, usually by means of some experiments or evaluations, comparing the performance against canonical systems when possible.

The specific methodology applied in each part of the dissertation can be summarized in:

- Design and implementation of a harmonic plus stochastic speech model including support of the RPS representation. For the analysis and interpretation of the RPS several tools have been developed: a specifically adapted pitch detection algorithm, CDP [23], polarity detection algorithms, inverse filtering tools, phasegram rendering tools, etc.
- Development of a parameterisation technique suitable for statistical modelling. The methodology here has been to evaluate different candidate parameter sets by means of an ASR application choosing those that maximise the accuracy of the system. The ASR used has been a canonical system (HTK) with a SpeechDat-like database in Basque language.
- In order to accomplish the objective of analysing how speech or speaker characteristics are reflected in the RPSs we have implemented and trained basic systems for speech recognition and speaker identification using the RPSs as parameters. Be-

sides evaluating the performance of the RPS based systems we have also compared it with the results of baseline MFCC systems, as well as combinations of both module and phase parameters. In all the cases the objective was not to evaluate the actual improvement RPS data can confer to state-of-the-art systems in these areas, but to test whether the RPSs carried information about these features of the speech signal.

- For the evaluation of the RPS application to the detection of synthetic impostors in speaker verification systems a module for synthetic speech detection was implemented, which works with the accepted signals of a standard GMM based speaker verification system. In this case we have compared the performance of the whole verification system with and without the synthetic detection module.
- Finally, for the evaluation of the perceptual importance of the phase manipulations in speech, we have designed a subjective quality evaluation aiming to highlight the subtle degradation that phase can introduce in speech signals. The evaluation procedure is based on the ITU-R BS 1116-1 recommendation for the subjective assessment of small impairments in audio [24], and was implemented in a web page. We have also designed and recorded a multispeaker and multilingual database to be used in this test: the utterances contain only voiced phonemes, and it consists of 10 sentences per speaker. There are speakers in 3 languages: 6 males and 6 females in Spanish and Basque languages, and 3 males and 3 females in German language.

4 Main Contributions of the Thesis

4.1 The Relative Phase Shift representation

The first contribution of the thesis is the Relative Phase Shift representation itself. From the instantaneous phases, calculated by means of the Fourier transform and the pitch, the proposed 'RPS transformation' allows calculating the RPS values in a pitch asynchronous way.

The new representation of the phases is useful when changing the duration and pitch of a speech signal by means of harmonic models. The RPSs allow calculating the phases for the modified signals in a straightforward way, similar to what it is done with the amplitudes. This way it is not necessary to do specific adjustments in the phases or limiting the analysis to specific pitch period points as other published methods require.

The RPS representation is more than a technique to ease phase manipulation. It shows many desirable properties, notably its direct relationship with the waveform of the signal and this allows direct visualization of the structure of the speech phases.

The relationship of the RPS with speech formants and source and vocal tract separation has also been analysed and described.

4.2 DCT-mel-RPS parameterization

The evident phase patterns that arise with the RPS phasegrams can be useful for humans but they are hardly usable by automatic applications. That is why we have developed a parameterization of the RPS values which is suitable for statistical modelling. This parameterization copes with some problematic characteristics of the RPS like wrapping, high dimensionality, variable number of values per frame at different frequencies, etc. The DCT-mel-RPS parameterization applies mel filtering and discrete cosine transform (DCT) to reduce the dimensionality of the original RPS values with the minimum loss of information. This parameterization has been tested with HMM based models in an ASR system and with GMM based models for speaker identification. In both experiments the parameterization has proven effective in this kind of statistical modelling.

In the experiments we have also proved that phase information is useful in these recognition systems either jointly with spectral module information (in the form of MFCC parameters) in the case of ASR, either just alone in the case of speaker identification. In both cases the inclusion of phase information can improve the results of traditional module only systems.

4.3 Polarity detector

One notable feature of the RPSs is the ability to distinguish the polarity of the signal. Inverting the polarity produces shifts of π radians in the RPSs values of the even components. Taking advantage of this fact we have proposed a polarity detection algorithm based on the comparison of the ripple of the RPS envelope along frequencies for every frame. This algorithm has been evaluated and compared with others reported in the literature and has obtained some excellent results even with short signals, outperforming the other algorithms.

4.4 Synthetic impostor detection

Phase information has been used to detect synthetic impostors in speaker verification systems with excellent results. This is the first time that the use of the signal phase has been proposed to discriminate the speech signal produced by synthesisers which imitate the voice of a specific person, and it opens a promising research line in an area where there were difficulties for this kind of impostors' detection with other methods.

The impostor detection system we have developed works as a separate stage after the verification system, trying to reject the synthetic impostors that have been wrongly accepted by the verification system. The synthetic speech detection (SSD) system models separately both the natural and the synthetic speech for every speaker, using the DCT-mel-RPS parameters, and classifies the claimant's speech accordingly.

4.5 Perceptual evaluation

One of the initial ideas for the thesis was to improve the treatment of the phases in the traditional speech synthesis systems where they are discarded or substantially modified. However, previously, we needed to know if caring about phases in synthesis was worth the effort, that is, if phase manipulations are perceived as impairments in the speech signal. To answer this question, we have evaluated speech signals with several phase transformations and compared them with the original one. This kind of experiments have been done before for other kind of audio signals (music, synthetic tones, etc.) but to our knowledge, there were no studies about the specific impact of the phase on the overall speech signal quality.

The results allow us to assure that human hearing is sensitive to the phase of the speech signal, and that major or less natural phase modifications are perceived as impairments. Nevertheless, concerning the convenience of adding the original phase information to the synthesis systems, the results are not categorical. Although the degradation produced by certain phase modifications is undeniable according to the results of the evaluation, it has to be noted that the test conditions were carefully designed to highlight the perception of these impairments, so their effect in a real system will probably be less perceptible than what our evaluation suggests.

In this sense it is worth noting the good results that the DCT-mel-RPS parameterization has obtained, with almost unnoticeable impairment in speech signals with phases calculated from this parameters.

5 Conclusions and opened research lines

As a result of this thesis, contributions were made in the speech phase representation and its application in different areas of speech processing, as has been detailed in the preceding section. Summing up, the thesis tries to highlight the usefulness of this new representation for the phases to be applied in various speech processing areas. The RPS representation removes the time dependency of the phase information, linking it to the waveform shape of the signal. These features make the RPS suitable to be used in applications where only spectral module information was used till now. We think that the experiments detailed in the thesis, though preliminary, can open the door for future and more extensive research to apply the RPS in these and other areas.

In this sense, one of the outcomes of this thesis is the opening of new research lines, which can be summarized in three blocks:

5.1 Improved RPS analysis

The calculation of the RPS can be improved to make it more robust to noisy signals, pitch errors or previous processing of the signal. This last aspect is crucial in order to apply the RPS analysis with telephone and coded speech signals.

5.2 Evaluation of the performance of the RPS in state of the art recognition systems

The experiments of the use of RPS information in ASR and speaker recognition systems detailed in the thesis were devised to check if the phase information was relevant for these tasks, and thus, classical recognition systems were employed. Since the results of these tests have proved the utility of the RPS, more research should be done to determine how the use of RPS data could improve the recognition rate in state of the art recognition systems, also using standard evaluation databases.

5.3 New application areas for the RPS

There are more areas of the speech processing domain where phase information can be applied. The results of the perception test clearly suggest the use of the DCT-mel-RPS model in a high quality statistic speech synthesis system. Also the low perceptual impact of the phases could be used to add information in the speech signal either for watermarking, or data hiding purposes.

6 Publications

The main results of this thesis have been published both in journals and conferences. The RPS representation has been described in a journal paper [25], and so has been the synthetic impostor detection [26]. Other results have been published in several important conferences: the polarity detection algorithm was described in [27] the application of the RPS in ASR systems has been reported in [28], speaker recognition results have been reported in [29], additional impostor detection experiments were also published in [30] and the evaluation of the perceptual importance of phases in speech signals has been published in [31].

Other research collaterally related with the thesis, namely the works around the pitch detection algorithm (CDP), have been published in [23] and [32]. This last paper received an award to one of the best student papers in the 5th JTH conference.

References

1. McAulay, R.J., Quatieri, T.F.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech and Signal Processing. 34, 744–754 (1986).
2. George, E.B., Smith, M.J.T.: Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. IEEE Transactions on Speech and Audio Processing. 5, 389–406 (1997).
3. Stylianou, Y.: Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification, PhD Thesis, (1996).
4. Erro, D.: Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models, PhD Thesis, (2008).
5. Quatieri, T.F., Mcaulay, R.J.: Shape invariant time-scale and pitch modification of speech. IEEE Transactions on Signal Processing. 40, 497–510 (1992).

6. Stylianou, Y.: Removing linear phase mismatches in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing. 9, 232–239 (2001).
7. O'Brien, D., Monaghan, A.: Shape Invariant Pitch and Time-Scale Modification of Speech Based on a Harmonic Model. Improvements in Speech Synthesis. pp. 64–75. Wiley Online Library (2002).
8. Liu, L., He, J., Palm, G.: Effects of phase on the perception of intervocalic stop consonants. Speech Communication. 22, 403–417 (1997).
9. Alsteris, L.D., Paliwal, K.K.: Evaluation of the modified group delay feature for isolated word recognition. Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005. pp. 715–718. IEEE (2005).
10. Pobloth, H., Kleijn, W.B.: On phase perception in speech. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings. pp. 11–14 (1999).
11. Kim, D.: On the perceptually irrelevant phase information in sinusoidal representation of speech. IEEE Transactions on Speech and Audio Processing. 9, 900–905 (2001).
12. Banno, H., Lu, J., Nakamura, S., Shikano, K., Kawahara, H.: Efficient representation of short-time phase based on group delay. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98. pp. 861–864. IEEE (1998).
13. Alsteris, L.D., Paliwal, K.K.: Short-time phase spectrum in speech processing: A review and some experimental results. Digital Signal Processing. 17, 578–616 (2007).
14. Murthy, H.A., Gadde, V.R.R.: The modified group delay function and its application to phoneme recognition. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP'03). pp. 2–5 (2003).
15. Padmanabhan, R., Parthasarathi, S., Murthy, H.A.: Robustness of phase based features for speaker recognition. Proc. Interspeech 2009. pp. 3–6 (2009).
16. Kua, J.M.K., Epps, J., Ambikairajah, E., Choi, E.: LS regularization of group delay features for speaker recognition. Proc. Interspeech 2009. pp. 2887–2890 (2009).
17. Hegde, R.M., Murthy, H.A., Gadde, V.R.R.: Application of the modified group delay function to speaker identification and discrimination. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. p. I–517–20. IEEE (2004).
18. Murty, K., Yegnanarayana, B.: Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters. 13, 52–55 (2005).
19. Zheng, N., Lee, T., Ching, P.C.: Integration of Complementary Acoustic Features for Speaker Recognition. IEEE Signal Processing Letters. 14, 181–184 (2007).
20. Wang, L., Minami, K., Yamamoto, K., Nakagawa, S.: Speaker identification by combining MFCC and phase information in noisy environments. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. pp. 4502–4505. IEEE (2010).
21. Masuko, T., Hitotsumatsu, T., Tokuda, K., Kobayashi, T.: On the security of HMM-based speaker verification systems against imposture using synthetic speech. Proceedings of the European Conference on Speech Communication and Technology. pp. 1223–1226. Citeseer (1999).
22. De Leon, P.L., Apsingekar, V.R., Pucher, M., Yamagishi, J.: Revisiting the security of speaker verification systems against imposture using synthetic speech. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1798–1801. IEEE (2010).
23. Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I.: Evaluation of pitch detection algorithms under real conditions. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. pp. 1057–1060 (2007).

24. ITU-R: ITU-R BS 1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, (1994).
25. Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sanchez, J.: Simple representation of signal phase for harmonic speech models. Electronics Letters. 45, 381–383 (2009).
26. De Leon, P.L., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I.: Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Transactions on Audio, Speech, and Language Processing. 20, 2280–2290 (2012).
27. Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., Navas, E.: Use of Harmonic Phase Information for Polarity Detection in Speech Signals. Proc. Interspeech 2009. 1075–1078 (2009).
28. Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D.: Using Harmonic Phase Information to Improve ASR Rate. Proc. Interspeech 2010. pp. 1185–1188 (2010).
29. Hernáez, I., Saratxaga, I., Sanchez, J., Navas, E., Luengo, I.: Use of The Harmonic Phase in Speaker Recognition. Proc. Interspeech 2011. pp. 2757–2760 (2011).
30. De Leon, P.L., Hernáez, I., Saratxaga, I., Pucher, M., Yamagishi, J.: Detection of synthetic speech for the problem of imposture. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 4844–4847 (2011).
31. Saratxaga, I., Hernaez, I., Pucher,, M., Navas, E., Sainz, I.: Perceptual Importance of the Phase Related Information in Speech. Proc. Interspeech 2012. ISCA, Portland, OR (2012).
32. Saratxaga, I., Luengo, I., Navas, E., Hernáez, I., Sánchez, J., Sainz, I.: Detección de pitch en condiciones adversas. Proc. V Jornadas en Tecnología del Habla. pp. 13–18 (2006).

Building task-oriented machine translation systems

Germán Sanchis-Trilles
Advisor: Francisco Casacuberta

Pattern Recognition and Human Language Technologies Group
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
gsanchis@dsic.upv.es

Abstract. This PhD dissertation, written by Germán Sanchis-Trilles and supervised by Francisco Casacuberta, was defended on June 20^{th}, 2012, at the Universitat Politècnica de València. The committee members were Enrique Vidal (UPV), Ismael García-Varea (UCM), Hermann Ney (RWTH-Aachen), José Oncina-Carratalá (UA), and Nicola Cancedda (XRCE). The qualification obtained was "sobresaliente cum laude". The main goal is to develop computer assisted translation and machine translation systems which present a more robust synergy with their potential users. Hence, the main purpose is to make current state-of-the-art systems more ergonomic, intuitive and efficient, so that the human expert feels more comfortable when using them. For doing this, different techniques are presented, focusing on improving the adaptability and response time of the underlying statistical machine translation systems, as well as a strategy aiming at enhancing human-machine interaction within an interactive machine translation setup. All of this with the ultimate purpose of filling in the existing gap between the state of the art in machine translation and the tools that are usually made available to the final human translators.

Keywords: Machine translation, adaptation, human-computer interaction

1 Introduction

In the last years, machine translation (MT) and computer assisted translation (CAT) is being increasingly embraced by human translators, who often find that post-editing the output of an MT system is an efficient way of producing high quality translations. Nevertheless, there are yet several problems which need to be dealt with before the usage of statistical machine translation (SMT) systems within CAT systems finds a more widespread usage. These problems concern mainly the time-efficiency, the adaptability, and the usability of the SMT systems. In this PhD dissertation, these three problems are confronted, yielding different degrees of success.

In the first place, state-of-the-art statistical machine translation (SMT) systems are often unable to yield real-time performance. This problem is even worse when the system has been trained on very large amounts of data, which is always desirable given that more data usually implies higher model coverage. When the amount of translation options and bilingual data made available to the system increases, translation throughput is necessarily affected, and model pruning strategies need to be applied with the

purpose of not having the human translator waiting too long for the system to produce its output, which would be on the one hand exasperating, and on the other hand economically inefficient. In this thesis, we focused on proposing a model pruning strategy which proves to be able to decrease system response time drastically, while keeping translation quality within state-of-the-art ranges.

Another topic tackled in this thesis is system adaptability. There is extensive work in SMT which proves that the translation quality produced by a typical machine translation system drops significantly when the text to be translated stems from a different topic than the data which has been used to train the system. In addition, different human translators may have different styles when translating a document, which implies that lexical choice or sentence length may be required to vary even when working within one single domain. Furthermore, from a user point of view it is mentally exhausting for a human translator to correct the same mistakes over and over again, while having the impression that those same mistakes will keep on appearing because the system is not learning from its own errors. For these reasons, system adaptability is unveiled as a key feature within a machine translation system that is setup within a human-machine collaborative framework. In the present thesis, two different model adaptation techniques are presented. The first one deals with the problem of language model adaptation, i.e., adapting the specific part of the translation system that controls word ordering and structure in the hypotheses produced. The second one deals with the adaptation of the translation model itself, which is the part of the translation system that will account for lexical choice and sentence length, among other features.

Lastly, usability of interactive machine translation systems is also a very important topic when attempting to build systems that are to be used by humans, whose expertise when using computers should not always be assumed. Hence, it is important to take special care when designing the interaction scheme, so that the human translator feels as comfortable as possible when using the translation interface. In this context, it is important to realise that the keyboard is not the only input device that the human user may use, but rather that richer interaction schemes might boost productivity. Nevertheless, it is also important to keep the interaction interface simple, so that the human expert is not overburdened. In this thesis, we propose a very simple and intuitive extension to the classical interactive machine translation interaction scheme, which takes into account the actions that the user performs before correcting a given hypothesis.

2 State of the art before this thesis

At the time of beginning this thesis, translation models were (and are still) often too large to fit into the memory of most table-top machines (let alone portable devices). This led to the widely-accepted practice of filtering the translation model according to the current test set to be translated. Even after doing so, state-of-the-art translation systems often took about 20 seconds to a whole minute when translating a single sentence, depending on its lenght. This fact led to the development of different test-independant translation model pruning strategies, some of which [1, 2] were published simultaneously to the first experiments reported in the present thesis. The technique presented in this thesis follows a similar direction to the one pointed out in [2], but instead of relying

on additional data for extracting usage statistics, it relies on the idea of re-estimating the current model parameters directly on the training data, thus discarding more parameters than reported in [2], while keeping translation quality mostly unaltered.

Regarding adaptation in SMT systems, there had been already some work in this direction. Specially concerning language model adaptation, the work presented in this thesis extends previous work already performed in [3, 4]. In these works, the training data is also partitioned in different ways with the purpose of extracting different sub-domains and building smaller topic-specific language models. The technique developed in this dissertation extends such work by presenting and comparing several different corpus subdivision strategies. In addition, the resulting language models are interpolated dynamically in translation time so as to optimise the perplexity of the input sentence, which had not been done as of yet. Regarding translation model adaptation, work performed before the beginning of this thesis mainly focused on examining different ways of combining the available training data [5, 6], did not confront the adaptation of the final translation model parameters directly [7], and the adaptation problem was seldom confronted from a purely statistical point of view. In this sense, the Bayesian adaptation strategy presented in this thesis confronts the adaptation of the model parameters directly, and has a solid statistical background.

Finally, user-machine interaction in interactive machine translation (IMT) had been remaining the same for the last years preceding the beginning of this thesis, i.e., the system waited for a keyboard interaction before performing any prediction [8]. With the introduction of the mouse as an alternative in the present thesis, it was proven that the classical IMT paradigm accepts numerous extensions, and since then there have been other works that have also attempted to extend such paradigm in different ways [9, 10].

3 Main contributions of the thesis

The main contributions of this thesis can be split into the three problems it tackles:

- A novel translation model pruning technique has been developed. Performing an unsmoothed re-estimation of the model parameters on the training data leads to much smaller models (about 3-5% of the original size), while translation quality remains mostly unaltered. The technique developed relies on the idea of translating the source side of the training data, selecting a set of best translations $G(\boldsymbol{x})$ for each input sentence, and then re-estimating the model parameters on those translations.
- Two novel adaptation techniques have been developed, one of them dealing with the adaptation of the language model, while the other one involves the adaptation of the translation model. The language model adaptation strategy first splits the training data so as to conform different smaller language models, which are then dynamically interpolated in translation time according to the input data. Results are not reported in the present paper, but are shown in the manuscript of the dissertation. The translation model adaptation strategy relies on the theoretical framework of Bayesian learning. Considering the model parameters as random variables whose prior distribution is influenced by the adaptation data leads to an effective adaptation framework. Since the Bayesian paradigm often implies computing an intractable integral, several different sampling strategies were implemented. In this

paper, only the best performing sampling methods are shown, namely Markov chain Monte-Carlo (MCMC) and a heuristic strategy (more details about this strategy in the document of the dissertation).

– The classical interactive machine translation framework is expanded by considering the mouse as an additional and valuable information source for the system. Two different pointer actions (PAs) were considered. The first one relies on the fact fact, before typing in a word, the user needs to position the cursor first. At this point, the system may already realise that the user intends to change the word located directly after the cursor (say y_i), so it is already able to provide a different translation completion, in which word y_i has been changed. This kind of pointer action was named anticipated proposal. In addition, the possibility of performing an explicit pointer action that explicitly refuses the system's proposal was also added. In this scenario, the user would perform a pointer action in front of a given word, thus explicitly asking the system to change it, together with the rest of the translation. This kind of pointer action was called partial refusal.

4 Corpus and methodology

With the purpose of giving soundness and replicability to the results reported in this thesis, all experiments were conducted by applying state-of-the-art SMT systems whenever possible, and using standard corpora that have been widely used in different SMT workshops. However, since the results presented correspond to a period of time that spans through five years $(2007 - 2012)$, some corpora were updated in the meantime.

Confidence intervals were computed for each one of the results reported, with the purpose of pointing out clearly which results are statistically significant, following the bootstrap strategy described in [11]. In the case of the experiments concerning adaptation, confidence intervals were computed by repeatedly extracting different random adaptation sets from the data available.

All the results presented in the present paper were obtained by using the open source SMT toolkit Moses [12], in its most standard setup, i.e., with direct and inverse translation models, lexical weights, a lexicalised re-ordering model [13] and a 5-gram language model estimated with the SRILM [14] toolkit.

Concerning the corpora used, all of the experiments reported on this paper were performed by using the Europarl corpus [15] as training data, in the partitions established in the WMT07 [16] and WMT10 [17] workshops on MT. The use of one or the other depended on the date in which the experiments were performed. Hence, the experiments concerning the parameter pruning strategy and effects of Bayesian adaptation are reported on the WMT10 partition, whereas the experiments conducted to assess the effectiveness of using pointer actions within an IMT framework are reported on the WMT07 partition. For the experiments concerning adaptation, the News-Commentary (NC) corpus was additionaly considered, also in the partition established in the WMT10 workshop. Unless stated otherwise, the log-linear weights of the translation model were always estimated on the `Devel.` set of the Europarl corpus. Different statistics of these corpora can be seen in Table 1.

System evaluation was performed by means of TER [18] and BLEU [19], whenever the experiments involved an SMT system, and WSR [8] for those involving an IMT

		De	En	Fr	En
WMT07	Sentences	751k		688k	
training	Running words	15.3M	16.1M	15.6M	13.8M
	Vocabulary size	195k	66k	80k	62k
WMT10	Sentences	1219k		1251k	
training	Running words	24.9M	26.1M	28.1M	25.6M
	Vocabulary size	255k	82k	101k	81k
	Sentences	2000		2000	
Devel.	Running words	55k	59k	67k	59k
	OoV 07/10	432/348	125/103	144/99	138/104
	Sentences	2000		2000	
Devtest	Running words	54k	58k	66k	58k
	OoV 07/10	377/310	127/111	139/114	133/112
	Sentences	86.9k		67.6k	
NC training	Running words	1.8M	1.8M	1.6M	1.4M
	Vocabulary size	86.7k	40.8k	43.3k	35.6k
	Sentences	2525		2525	
NC 09 test	Running words	62.7k	65.6k	72.6k	65.6k
	OoV NC/10	3629/2410	1853/1247	2478/1446	2035/1247

Table 1. Characteristics of Europarl and News-Commentary, for each of the sub-corpora. OoV stands for "Out of Vocabulary" words with respect to the WMT07 training data (07), the WMT10 training data (10), or the NC training data (NC). Devel. stands for Development, k for thousands of elements and M for millions of elements.

system. TER is an error metric (i.e., the lower the better) that measures the minimum amount of edits required to transform the system's hypothesis into the reference sentence. BLEU is a precision metric (i.e., the higher the better) that measures n-gram precision, with a penalty for sentences that are too short. WSR measures the ratio of words that a human translator would need to type before achieving the sentence he has in mind (i.e., the lower the better).

5 Experimental results

5.1 Model size reduction

For assessing the effectiveness of the model pruning strategy described in Section 3, a baseline model was estimated by means of Moses, as described in the previous section, by using the WMT10 training data for training the initial models and the `Devtest` set as evaluation set. Results for different language pairs are available in the document of the dissertation, but in this paper we only report results on French–English. Different sizes of $G(\boldsymbol{x})$ (see Section 3 were considered, both within a SMT and an IMT setting.

The results for the SMT case are shown in Figure 1, whereas the results for IMT are shown in Figure 2. It can be seen that the reduction strategy implemented has a very important impact in system speed, while mantaining translation quality almost unaltered in the case of SMT. It is worth noting that, for $|G(\boldsymbol{x})| = 1$, only about 3% of the original paremeters are retained.

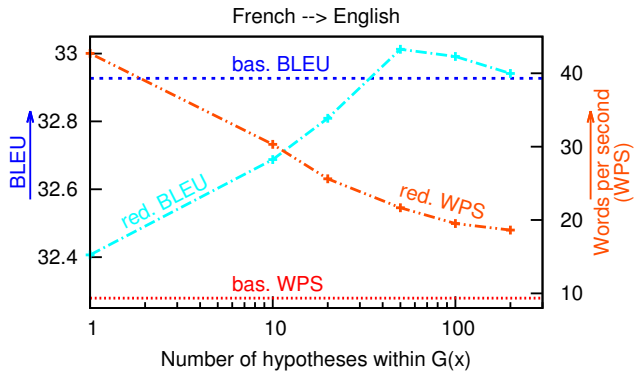


Fig. 1. Model pruning within in SMT. `bas.` stands for baseline, and `red.` for the pruned system.

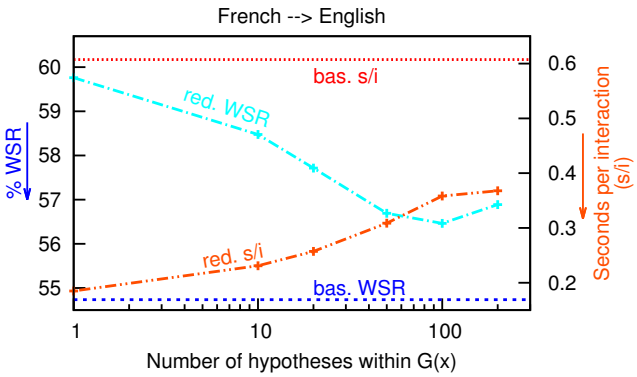


Fig. 2. Effect of model pruning within an IMT scenario.

In the case of IMT, the results suggest that a certain trade-off is required. Although the speed increase is also very significative, the increase in the amount of interactions required by the human translator may not be acceptable for small sizes of $G(\boldsymbol{x})$. However, it must be considered that response time in IMT is critical specially when the suffix to be produced is long, which implies that the technique implemented might be useful in such cases, but not when the suffix to be produced is short.

5.2 Bayesian translation model adaptation

The effectiveness of the Bayesian translation model adaptation strategy (Section 3) was tested by training a baseline SMT system on the German–English WMT10 training data. Then, the NC training data was used for randomly extracting adaptation sets of different sizes, with the purpose of analysing the behaviour when increasing the amount of adaptation data, and the adapted systems were evaluated on the NC09 test set. The results are shown in Figure 3, for the two sampling strategies that yielded the best results. As shown, the two Bayesian adaptation strategies perform better than the complete re-estimation of the log-linear weights. Although not shown here, confidence intervals

were also found to be much smaller for the two Bayesian approaches than for MERT (although not for MERT+), implying more relyiable results.

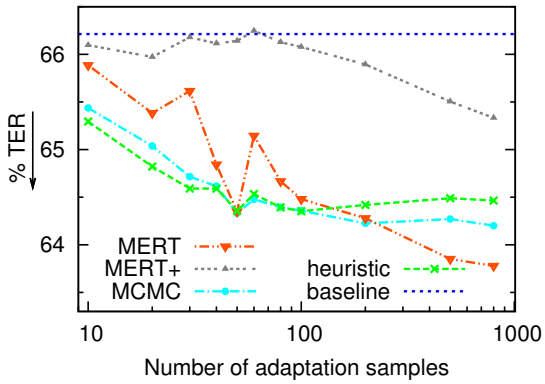


Fig. 3. Translation quality comparison between re-estimating $\boldsymbol{\lambda}$ from scratch and Bayesian adaptation for different sampling strategies. MERT stands for re-estimating the log-linear weights from scratch on the adaptation data, and MERT+ for re-estimating them on the concatenation of adaptation data and development set.

Additionally, a temporal analysis was also conducted, and is shown in Figure 4. As shown, the two Bayesian approaches perform about half an order of magnitude faster than the MERT approach, and two orders of magnitude faster than the MERT+ approach. When computing the time required by each system, all of the steps were considered (e.g., computing the sentence-level TER counts for the Bayesian approaches). We find that these results are very encouraging, since when a user is sitting waiting for the translation to be produced, computational time is critical. Although not presented here for space reasons, online adaptation experiments are also shown in the document of the dissertation, reporting encouraging results.

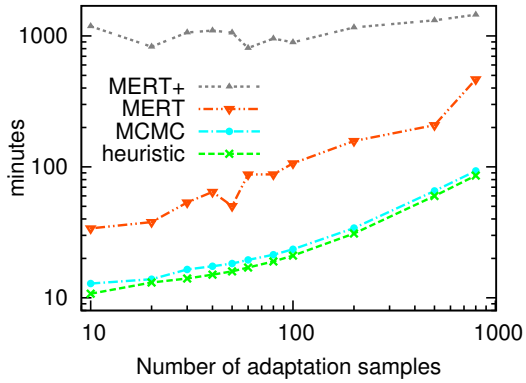


Fig. 4. Temporal comparison between re-estimating $\boldsymbol{\lambda}$ from scratch and Bayesian adaptation for different sampling strategies.

5.3 Enriching user-machine interaction

For assessing the effectiveness of the novel interaction scheme described in Section 3, an IMT scenario was simmulated by considering that the sentence the user would want to obtain is the reference present in the test set. The French–English WMT07 training data was used to train the SMT translation models, and system performance was evaluated on the `Devtest` set (Figure 5). Experiments concerning other language pairs lead to similar results, and are omitted here for space reasons.

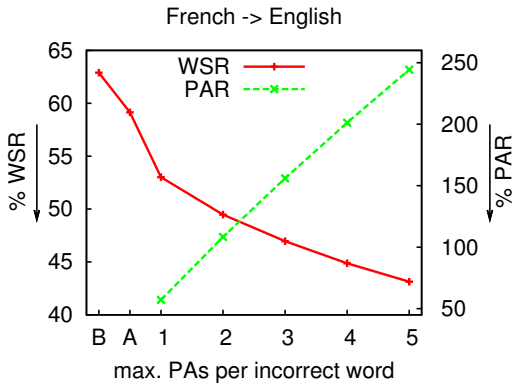


Fig. 5. IMT system performance when introducing pointer actions. `B` stands for baseline (i.e., the classical interaction scheme), `A` for teh system that incorporates the anticipated proposal, and considering up to n partial refusal pointer actions. `PAR` is the ratio of pointer actions performed.

As shown in Figure 5, considering the anticipated proposal pointer action already leads to a reduction of about 5% in the amount of words typed by a human translator. Considering only one partial refusal pointer actions leads to a further drop of more than 5%. However, considering further partial refusal pointer actions has less effect, and the benefit tends to fade for larger amounts of pointer actions. Hence, it seems quite unlikely that someone would perform more than 2 or 3 pointer actions. However, by just performing 3, the potential user would already spare typing in about 10% of words.

6 Conclusions

As a result of this thesis, several contributions to the human-machine interaction framework were made. The response time of state-of-the-art SMT and IMT systems was improved, without any loss of translation quality in the former case, requiring a certain compromise between time efficiency and human effort in the latter. System adaptability was also improved in two key aspects of a typical SMT system, such as the translation model and the language model. Results in the first case achieved very promissing improvements with respect to the baseline system. In addition, the classical user-machine interface of an IMT system was extended with very promissing results. All these strategies and ideas were tested within state-of-the-art SMT and IMT systems, trained on publicly available corpora which have been used in recent SMT workshops, and yielded satisfactory results, as shown by the experiments reported.

7 Publications

Preliminary work regarding the parameter pruning strategy described in Section 3 was published in an international workshop [20] and an international conference [21]. A more refined version of such work, and re-formulated as a parameter re-estimation technique, was presented later on in an international conference [22]. The language model adaptation technique lead to two publications in an international workshop [23, 24], and a publication in an international conference [25]. Related work was also published in another international workshop [26]. The Bayesian translation model adaptation strategy lead to one publication in an international workshop [27] and another publication in an international conference [28]. The online variant was published in the Pattern Recognition Journal [29], and the batch variant was used within an SMT system presented in an international MT competition [30]. Finally, the interaction framework described in Section 3 was accepted for publication in an international workshop [31] and in an international conference [32]. Four of these publications [32, 28, 27, 26] are ranked as high impact by the Computer Research and Education Association of Australasia (CORE).

Acknowledgements

Work partially supported by the EU 7th Framework Programme (FP7/2007-2013) under grant Nr. 287576 (CasMaCat), by the EC (FEDER/FSE) and the Spanish MICINN under projects MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) and iTrans2 (TIN2009-14511), and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

1. Johnson, H., Martin, J., Foster, G., Kuhn, R.: Improving translation quality by discarding most of the phrasetable. In: Proc. of EMNLP/CoNLL. (June 2007) 967–975
2. Eck, M., Vogel, S., Waibel, A.: Translation model pruning via usage statistics for statistical machine translation. In: Proc. of NAACL/HLT. (April 2007) 21–24
3. Zhao, B., Eck, M., Vogel, S.: Language model adaptation for statistical machine translation with structured query models. In: Proc. of COLING. (August 2004) 411–417
4. Lü, Y., Huang, J., Liu, Q.: Improving statistical machine translation performance by training data selection and optimization. In: Proc. of EMNLP/CoNLL. (June 2007) 343–350
5. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proc. of the ACL second workhop on SMT. (June 2007) 224–227
6. Bertoldi, N., Federico, M.: Domain adaptation in statistical machine translation with monolingual resources. In: Proc. of the EACL fourth workshop on SMT. (March 2009) 182–189
7. Civera, J., Juan, A.: Domain adaptation in statistical machine translation with mixture modelling. In: Proc. of the ACL second workhop on SMT. (June 2007) 177–180
8. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E.: Statistical approaches to computer-assisted translation. Computational Linguistics **35**(1) (2009) 3–28
9. Alabau, V., Ortiz-Martnez, D., Sanchis, A., Casacuberta, F.: Multimodal interactive machine translation. In: Proc. of ICMI/MLMI. (2010) DOI: 10.1145/1891903.1891960.
10. Alabau, V., et. al: On multimodal interactive machine translation using speech recognition. In: Proc of ICMI. (2011) 129–136

11. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proc. of EMNLP. (July 2004) 388–395
12. Koehn, P., et. al: Moses: open source toolkit for statistical machine translation. In: Proc. of ACL: Demo and Poster Sessions. (June 2007) 177–180
13. Koehn, P., et. al: Edinburgh system description for the 2005 iwslt speech translation evaluation. In: Proc. of IWSLT. (October 2005)
14. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of the Int. Conf. on Spoken Language Processing. (September 2002) 901–904
15. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the MT Summit X. (September 2005) 79–86
16. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: Proc. of the ACL workshop on SMT. (June 2007) 136–158
17. Callison-Burch, C., et al.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proc. of the ACL workshop on SMT and Metrics MATR. (July 2010) 17–53
18. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proc. of AMTA. (August 2006) 223–231
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Technical Report RC22176 (W0109-022). (2001)
20. Sanchis-Trilles, G., Casacuberta, F.: Increasing translation speed in phrase-based models via suboptimal segmentation. In: Proc. of PRIS. (June 2008)
21. González, J., Sanchis-Trilles, G., Casacuberta, F.: Learning finite state transducers using bilingual phrases. In: Proc. of CICLing. (February 2008)
22. Sanchis-Trilles, G., Ortiz-Martínez, D., González-Rubio, J., González, J.: Bilingual segmentation for phrasetable pruning in statistical machine translation. In: Proc. of EAMT. (May 2011)
23. Sanchis-Trilles, G., Cettolo, M., Bertoldi, N., Federico, M.: Online language model adaptation for spoken dialog translation. In: Proc. of IWSLT. (December 1 – 2 2009) 160–167
24. Bertoldi, N., Bisazza, A., Cettolo, M., Sanchis-Trilles, G., Federico, M.: Fbk @ iwslt 2009. In: Proc of IWSLT. (December 2009) 37–44
25. Sanchis-Trilles, G., Cettolo, M.: Online language model adaptation via n-gram mixtures for statistical machine translation. In: Proc. of EAMT. (May 2010)
26. Andrs-Ferrer, J., Sanchis-Trilles, G., Casacuberta, F.: Similarity word-sequence kernels for sentence clustering. In: Proc. of S+SSPR. (August 2010)
27. Sanchis-Trilles, G., Casacuberta, F.: Bayesian adaptation for statistical machine translation. In: Proc. of S+SSPR. (August 2010)
28. Sanchis-Trilles, G., Casacuberta, F.: Log-linear weight optimisation via bayesian adaptation in statistical machine translation. In: Proc. of COLING. (August 2010)
29. Martínez-Gómez, P., Sanchis-Trilles, G., Casacuberta, F.: Online adaptation strategies for statistical machine translation in post-editing scenarios. Pattern Recognition **45**(9) (2012) 3193–3203
30. Gascó, G., Alabau, V., Andrés-Ferrer, J., González-Rubio, J., Rocha, M.A., Sanchis-Trilles, G., Casacuberta, F., González, J., Sánchez, J.A.: Iti-upv system description for iwslt 2010. In: Proc. of IWSLT. (December 2010)
31. Sanchis-Trilles, G., González, M., Casacuberta, F., Vidal, E., Civera, J.: Introducing additional input information into imt systems. In: Proc. of MLMI. LNCS. Volume 5237. (September 2008) 284–295
32. Sanchis-Trilles, G., Ortiz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., Hoang, H.: Improving interactive machine translation via mouse actions. In: Proc. of EMNLP. (October 2008)

Robust Diarization for Speaker Characterization

Carlos Vaquero[1] and Alfonso Ortega[2]

[1] Agnitio S.L.,
cvaquero@agnitio-corp.com
[2] Aragon Institute for Engineering Research(I3A)
ortega@unizar.es

Abstract. There are many applications related to speaker characterization, specially in telephone environments, where large datasets are available but not directly useful since there are several speakers involved in every recording. In this situation, the use of accurate speaker diarization systems is mandatory. In this work we propose a new approach for speaker diarization with a novel variability compensation technique that shows a significant improvement in the task of diarization of two-speaker telephone conversations. However, even with very accurate speaker diarization systems, we can expect to find some recordings with low diarization accuracy. The use of these recordings can reduce the accuracy of any speaker characterization technology. To overcome this problem, a set of confidence measures to assess the quality of a hypothetical diarization output is presented. We show that these confidence measures enable us to retrieve most of the desired recordings from a given dataset, discarding those recordings that degrade the overall accuracy of an application that make use of speaker characterization technologies.

Keywords: Speaker Diarization, Quality Assessment, Intra-session Variability Compensation, Speaker Clustering

1 Introduction

The task of speaker characterization, which aims at describing the particular and distinctive peculiarities of a person's speech, is essential for several speech based technologies and applications, such as voice based biometrics or speech recognition. Speaker characterization techniques require large datasets with speaker labels to operate. However in several environments, these datasets are not directly useful for speaker characterization, since every recording in the dataset contains several speakers, and there are no labels available indicating when every speaker is speaking. The solution to this problem is the use of speaker diarization, which aims at answering the question *Who spoke when?*.

This document summarizes the work developed in the PhD thesis with title *Robust Diarization for Speaker Characterization* [1]. This thesis focuses on providing robustness to speaker diarization for real life speaker characterization applications. For this purpose two complementary objectives are pursued: First,

the development of accurate speaker diarization systems is desired in order to ensure that speaker characterization applications will operate correctly when they make use of recordings containing more than a single speaker. Second, quality assessment strategies for speaker diarization are desired in order to detect those recordings that will be reliable for speaker characterization.

To achieve the first objective, a new approach for speaker diarization based on the most recent innovations in the field of speaker recognition is proposed, including a novel variability compensation strategy for speaker diarization.

The second objective is achieved carrying out a study on quality assessment strategies for speaker diarization. Several confidence measures and a methodology to detect recordings with reliable diarization hypotheses are proposed. This methodology enables us to retrieve a subset of reliable recordings, ensuring that a speaker characterization application will not obtain significant degradation due to speaker diarization errors when the retrieved subset is considered.

Accurate speaker diarization systems have been an aim for the research community in the recent years, in part motivated by the NIST Rich Transcription (RT) Evaluations. Currently, the best performing systems proposed in these evaluations are mostly based on Agglomerative Hierarchical Clustering (AHC) algorithms using Bayesian Information Criterion (BIC) [2] as distance metric [3].

On the other hand, the recent advances in the field of speaker recognition have motivated new approaches for speaker diarization. Most of them are based on the use of Factor Analysis (FA) models [4] and include the use of eigenvoice modeling [5], [6], soft-clustering [7], or Variational Bayes methods [8]. All these approaches have shown better performance than the traditional BIC based AHC, but only in two-speaker telephone conversations.

However, there is little work in the field of assessing the quality of diarization hypotheses. The most relevant work in this field that can be found in the literature has been obtained as a result of this thesis: In [6], a small set of confidence measures is proposed aiming at predicting the performance of a speaker diarization system for telephone conversations. In [9], it is shown that these confidence measures enable us to improve the performance when used to select the best diarization hypothesis among several hypotheses available for every recording.

This document is organized as follows: In Section 2 the proposed approach for speaker diarization is described, and in Section 3 we describe a quality assessment methodology for speaker diarization. These techniques are evaluated in Section 4 to summarize the main conclusions in Section 5.

2 Speaker Diarization

The speaker diarization system presented in this thesis makes use of an eigenvoice model [10] in order to obtain a compact representation of a single speaker. This compact representation is a low dimension vector y, whose components are known as speaker factors. Such a representation has the advantage that, compact as it is, does not need much data to be estimated.

In order to obtain this compact representation a Factor Analysis (FA) model [4] is built. Assuming that a set of feature vectors $X = \{x(1), x(2), ...x(F)\}$ of dimension D has been extracted from a recording or set of recordings that belongs to a single speaker s, and that a GMM Universal Background Model (UBM) is available, we can build a model for speaker s as follows:

$$M_s = M_{UBM} + Vy_s, \tag{1}$$

where M_s is the speaker s GMM supervector of dimension CD, y_s is a set of R speaker factors that represents the speaker s, and V is a $CD \times R$ low rank eigenvoice matrix that models inter-speaker variability, capturing those directions of maximum variability among different speakers. To fulfill the FA model assumptions, the speaker factor vector must follow a Normal Standard distribution $N(y|0, I)$ a priori.

To perform speaker segmentation using this FA model, given a recording that may contain speech from different speakers, we estimate the posterior distribution of $y(t)$ for small overlapped segments $t = 1, .., T$, according to the FA model presented in [4]. Only one speaker is assumed to be active in every segment t, and that speaker will be represented as a point estimate given by the mean of the posterior of $y(t)$, $m_{y(t)}$. Therefore, $m_{y(t)}$ will be a compact representation of the speaker present in every segment t. This way, the problem of speaker segmentation reduces to a clustering problem, where the speaker factors associated to the same speaker should be clustered together. Since we know that for a given speaker s the posterior distribution of y_s is normal, the problem of two-speaker segmentation reduces to finding the two Gaussian models that generated the obtained stream of speaker factors $Y = m_{y(1)}, ..., m_{y(T)}$.

2.1 System architecture

We describe the speaker diarization system for two-speaker telephone conversations proposed in the thesis [1]. In order to perform speaker diarization given a recording containing two speakers, we follow the steps shown in Fig. 1 which are described below:

Front End: The first stage extracts a sequence of MFCC feature vectors from the input audio signal, and then it computes the sequence of speaker factors. Several configurations for the Front End have been evaluated, and the best configuration considers 19 MFCC including C0 plus Δ features with no compensation. Speaker factor vectors of dimension $R = 100$ are estimated every 10 ms considering a window of 1 second length.

Variability Compensation: A study of the variability sources that affect a speaker diarization system has been carried out in the thesis [1], concluding that the compensation of the variability that a speaker shows during a recording session can improve the accuracy of the diarization system. This intra-session variability is compensated considering LDA and WCCN as presented in [11].

Initial Clustering: The speaker factors processed as previously explained enable easy speaker separation. The system applies PCA on the compensated

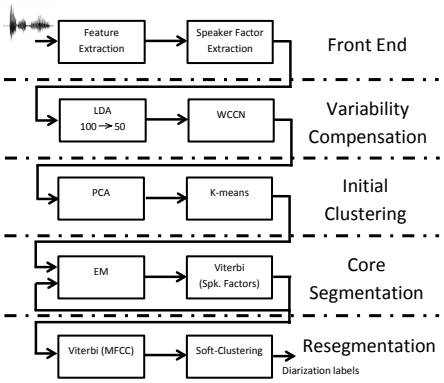


Fig. 1. *Block diagram of the proposed diarization system*

speaker factors to obtain the best direction to separate the two speakers. Two clusters are obtained using uniquely this direction, and then refined using K-means on the fully dimension speaker factor vectors.

Core Segmentation: The clusters previously obtained serve to build an initial two-component GMM for the whole recording. Then a two stage iterative process is applied until convergence. First the two Gaussians that best fit the stream of compensated speaker factors are estimated. Each one of these Gaussians is assigned to a single speaker. Second, an HMM with a left-to-right sequence of 10 tied-states, is build for every speaker, and the compensated speaker factor vectors are reassigned to the speakers using Viterbi decoding.

Resegmentation: The labels obtained from the Core Segmentation are refined by means of Viterbi resegmentation using MFCC as features. In this case we model every speaker using a left-to-right sequence of 10 tied-states sharing a 32-component GMM as observation distribution. The use of MFCC for resegmentation is motivated by the fact that they provide a much higher temporal resolution than the speaker factors. After this Viterbi resegmentation we retrain the GMM speaker models and run a soft-clustering pass [7].

3 Quality assessment for Speaker Diarization

In the previous sections, we have presented a state-of-the-art speaker diarization system for two-speaker telephone conversations that is expected to obtain a very good diarization accuracy for most of the recordings of a dataset. However, this system or any other diarization system will be much more useful if for every diarization hypothesis, a set of measures to assess the quality of such segmentation could be provided, in order to retrieve those recordings correctly diarized. These measures are then confidence measures for speaker diarization, and enable us to assess the quality of a diarization hypothesis.

3.1 Confidence measures

In this work a total of nine confidence measures for speaker diarrization have been analyzed, including four separability indicators and five well-conditioned data indicators. The separability indicators are: BIC, Kullback-Leibler (KL) divergence, cosine distance with intra-session variability compensation and PLDA speaker verification score. These measures can be considered as distance metrics between two speaker models built from a diarization hypothesis. We expect the distance between both models to be high if the segmentation is correct, i.e. the speaker models are different and pure. The well-conditioned data indicators are the UBM likelihood, the number of iterations that the Core Segmentation takes to converge, the Skewness and the Kurtosis of the speaker factors, and the Normalized Eigenvalue Spread of the speaker factors. These measures try to determine whether or not the input data to the speaker diarization system are well conditioned and whether or not the system assumptions are fulfilled.

Five confidence measures were discarded because they did not show correlation with the Diarization Error Rate (DER) or were highly correlated with other confidence measures and thus did not provide additional information. The selected confidence measures are described next:

BIC: Given two sequences of acoustic feature vectors X_1, X_2 segregated by the segmentation system, we compute the BIC for two hypotheses: Each sequence belongs to a different speaker or both sequences belong to the same speaker (joint hypothesis). The confidence measure is the difference between both BIC values. To avoid adjusting BIC penalty parameters, we force the models for both hypotheses to have the same complexity.

i-vector speaker verification score: Given two segments, we want to determine whether every segment was uttered by a single speaker. Otherwise, at least one speaker will be present in both segments. Therefore, this problem is related to speaker verification, in the sense that we want to detect whether the same speaker is present in both segments, and a speaker verification system could give a measure of how probable are both segments to be uttered by the same speaker. Such measure can be used as confidence measure for this task. As speaker verification system we use an i-vector system [11]. In this case, we replace the i-vectors with the speaker factors obtained directly with the same eigenvoice matrix V used for segmentation, with $R = 100$, and we consider all available data to compute the speaker factor vector for every speaker. Intra-session variability is compensated using LDA+WCCN.

Probabilistic LDA (PLDA) speaker verification score: Recently, a new approach for speaker verification when working with i-vectors has been proposed in [12]. Such approach is similar to that presented in [11] but instead of using LDA to find the subspace that best discriminate among different speakers, it uses PLDA to model such subspace, and optionally an inter-session variability subspace. The main advantage of this approach against traditional LDA is that, since it is probabilistic, it provides a way to compute the likelihood of a speaker model, and thus the score can be computed in a similar way to BIC.

Normalized Eigenvalue Spread of the speaker factors: The proposed speaker diarization system works under the assumption that the speaker factor vectors for a given speaker follow a Gaussian distribution whose covariance is close to the identity. This assumption is critical for the initial clustering performed by the diarization system, since it makes use of PCA and K-means. If this assumption is not fulfilled, we can expect certain degradation in the accuracy of the speaker diarization system. So an indicator of how close our speaker models are to fulfill this assumption is a good candidate for confidence measure. We propose a normalized eigenvalue spread estimation defined as:

$$C_{spread} = log \frac{\frac{max(\lambda_{1,2})}{median(\lambda_{1,2})}}{\frac{max(\lambda_1)}{median(\lambda_1)} \frac{max(\lambda_2)}{median(\lambda_2)}} \quad (2)$$

where λ_s are the eigenvalues of the covariance matrix of speaker s estimated using y_s. In all eigenvalue spread computation the median of the eigenvalues have been used in the denominator rather than the minimum, since the minimum may be noisier.

3.2 Detecting Correctly Segmented Recordings

Obtaining a reliable confidence measure can be useful to predict the performance of the segmentation system for a segmentation hypothesis, so that a given application can decide how to deal with the current recording. This usually means that the confidence measure is compared to a threshold or set of thresholds in order to classify the recording into different classes that will be processed differently. Then, given a dataset, a partition is created according to the predicted accuracy of the diarization system so that the application can deal properly with every class in the partition. Therefore, the partition will be application dependent. For example, a semi-supervised segmentation system can be built, so that the user only needs to check the segmentation hypotheses for a small subset of the whole dataset. This subset will be composed of those recordings that the quality assessment system labels as unreliable.

In this study we limit our problem to two classes, assuming that our application will use only the subset of correctly segmented recordings and discard the rest of the dataset. This way, we want to solve the detection problem where only those segmentation outputs obtaining a DER below a threshold ϵ are desired.

In order to detect those recordings correctly segmented, we can train a weighted linear logistic regression model fusing the four proposed confidence measures. This model will give as output the certainty of a hypothetical segmentation for a given recording to be correct.

The weighted linear logistic regression model is trained to maximize a figure of merit introduced in the thesis [1], which measures the usefulness of a dataset that must be processed with a diarization system. This figure of merit takes into account both the reliability of the dataset (the accuracy of the diarization system) and its representativeness (portion of the dataset that is kept).

4 Evaluation

The techniques proposed in this thesis are designed to increase the accuracy of speaker diarization, but the final purpose is to obtain useful datasets for speaker characterization. In this section we evaluate the proposed diarization system and the quality assessment methodology in a diarization task, but we also consider a speaker verification task to analyze the impact of these methods in a speaker characterization application. Other applications can be considered, in fact, in this work [1], the proposed techniques show great improvement in a speaker clustering task and also as initialization to improve the diarization accuracy.

4.1 Experimental setup

As speaker diarization task, we consider the diarization of the NIST SRE 2008 summed condition. This condition includes 2213 two-speaker telephone conversations. The accuracy of the diarization system is measured in terms of DER and also in terms of rate of correctly segmented recordings. It is assumed that a diarization hypothesis is considered to be correct when the DER is below $\epsilon = 10\%$. This threshold is suitable for a speaker verification task as is shown in this thesis and in [7]. We also analyze the accuracy of the quality assessment approach for the detection of correctly segmented recordings, evaluating the accuracy of the diarization system in both the selected and the rejected subsets.

Finally, we analyze the impact of the proposed techniques in a speaker verification task that makes use of the NIST SRE 2008 summed dataset. We consider an extended NIST SRE 2008 short2-summed condition where all possible trials are performed. We consider a state-of-the-art PLDA speaker verification system. Results are measured in terms of EER and minimum of the Detection Cost Function (minDCF) as defined by NIST [16].

The results obtained with the proposed methodologies are compared to those obtained with those obtained using a traditional BIC AHC based system.

4.2 Diarization results

Table 1. Accuracy of the proposed diarization techniques on the NIST SRE 2008 summed dataset

System	DER	Rate $DER < 10\%$	% of dataset
BIC AHC	5.21%	85.02%	100.00%
Spk Fact	1.77%	92.91%	100.00%
Spk Fact + Comp	1.31%	95.26%	100.00%
Spk Fact + Comp + Q.A. selected subset	0.86%	99.36%	97.38%
Spk Fact + Comp + Q.A. rejected subset	19.96%	29.39%	2.62%

Table 1 shows the accuracy of the proposed diarization system and quality assessment method. Firstly, note that the proposed speaker factor diarization

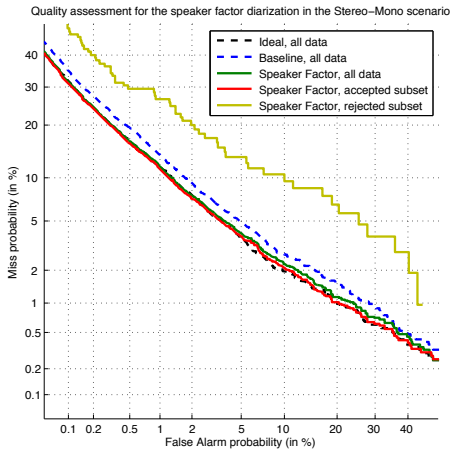


Fig. 2. *DET curves for the extended short2-summed condition.*

system (*Spk Fact*) clearly outperforms the baseline BIC AHC, obtaining a DER of 1.77%, and a total of 92.91% of the database with a DER below $\epsilon = 10\%$. Introducing intra-session variability compensation (*Comp*) reduces the DER to 1.31%, and to increase the rate of correctly segmented recordings to 95.26%.

Finally, the Quality Assessment method (*Q.A.*) enables us to retrieve the 97.38% of the dataset that obtains an overall DER as low as 0.86%. Almost all the recordings (99.36%) in the subset selected as correctly segmented by the quality assessment method are actually correctly segmented. The subset rejected and labeled as incorrectly segmented is only the 2.62% of the dataset, and has a DER as high as 19.96%. Only 29.39% of the recordings in this subset are correctly segmented.

4.3 Speaker Verification results

Figure 2 compares the DET curves obtained considering an ideal or perfect diarization system, the baseline BIC AHC system, the proposed diarization system with intra-session variability compensation, and the selected and rejected subsets obtained after quality assessment, for the speaker verification task.

Table 2. EER and minDCF for the extended short2-summed condition.

System	EER (degradation)	minDCF (degradation)
Ideal	4.23% (0.00%)	0.2102 (0.00%)
BIC AHC	4.94% (16.78%)	0.2334 (11.04%)
Spk Fact + Comp	4.39% (3.78%)	0.2097 (-0.24%)
Spk Fact + Comp + Q.A. selected subset	4.26% (0.62%)	0.2056 (-2.19%)
Spk Fact + Comp + Q.A. rejected subset	9.44% (122.99%)	0.3429 (63.16%)

The use of the proposed system obtains better speaker verification results than the use of BIC AHC. In fact, the proposed system obtains an accuracy very close to an ideal diarization system. Note also that the quality assessment strategy is working as expected: the selected subset obtains similar results to those obtained with an ideal diarization system, while the rejected subset obtains very poor results. These effects can also be observed in Table 2 that shows the EER and minDCF and the relative degradation when compared to an ideal diarization system for the systems under analysis. Note that the results for the selected subset are not much better than those for the whole dataset since the selected subset comprises the 97.38% of the dataset. We can expect to obtain higher improvement when the quality assessment strategy faces a harder dataset with fewer correctly segmented recordings.

5 Conclusions

In this work we have presented a speaker factor based diarization system with a set of improvements that enables us to obtain competitive results when compared to state-of-the-art speaker diarization systems. Among these improvements, the most relevant contribution of this thesis is the use of intra-session variability compensation for speaker diarization. This has shown an important improvement in the accuracy of speaker diarization, reducing the DER from 1.77% to 1.31% in the NIST SRE 2008 summed dataset.

But the main contribution of this work is the quality assessment methodology for speaker diarization. The concept of confidence measures and quality assessment has been studied in other fields of speech technologies, but never in the field of speaker diarization. In this work, we propose a set of confidence measures along with a weighted logistic regression model in order to determine whether or not a recording has been correctly segmented and thus it is reliable for its use in a speaker characterization application. We have validated this in a speaker verification task, considering the short2-summed condition, showing that the proposed quality assessment methodology enables us to retrieve the 97.38% of the summed dataset with a DER as low as 0.86%. The discarded subset is small and obtains a DER as high as 19.96%, and a degradation in the speaker verification task of 122.99% in EER and 63.16% in minDCF. In a real application, the discarded subset could be retrieved by manually inspection.

As proof of the quality of this work, we can enumerate five contributions to international conferences directly related to this thesis: The first one analyzes the vulnerability of FA to unseen variability [13]. The second one proposes the speaker factor based diarization system and a first approach for quality assessment [6]. The third one proposes a hybrid approach for online speaker diarization [14]. The fourth one studies different strategies for variability compensation in speaker diarization [9]. The last one analyzes the task of speaker clustering in large datasets composed of two-speaker conversations [15]. In addition, a journal article including the complete methodology for quality assessment presented in this thesis has been submitted [17].

But the most important proof of the quality of this work is the technology transfer to the industry. In fact, all the work in this thesis is developed with the goal of being useful for real life in the future. Currently, the company Agnitio S.L. is providing diarization solutions for speaker verification based on the new approaches presented in this thesis.

References

1. C. Vaquero.: Robust Diarization for Speaker Characterization. PhD Thesis, University of Zaragoza, 2011. Available at: `http://zaguan.unizar.es/record/7065/files/TESIS-2012-038.pdf`.
2. S. S. Chen and P. S. Gopalakrishnan.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. In: IEEE ICASSP 1998, pp. 127–132, 1998.
3. D. Reynolds and P. Torres-Carrasquillo.: Approaches and Applications of Audio Diarization. In: IEEE ICASSP 2005, pp. 953–956, Philadelphia, PA, 2005.
4. P. Kenny et al.: A Study of Interspeaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing, 16(5), 980–988, July 2008.
5. F. Castaldo et al.: Stream-based Speaker Segmentation Using Speaker Factors and Eigenvoices. In: IEEE ICASSP 2008, pp. 4133–4136, Las Vegas, Nevada, 2008.
6. C. Vaquero et al.: Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification. In: Interspeech 2010, pp. 2310–2313, Makuhari, Tokyo, 2010.
7. D. Reynolds, P. Kenny and F. Castaldo. A Study of New Approaches to Speaker Diarization. In: Interspeech 2009, Brighton, UK, 2009.
8. P. Kenny, D. Reynolds, and F. Castaldo. Diarization of Telephone Conversations using Factor Analysis.: IEEE Journal on Selected Topics in Signal Processing, 4(6), 1059–1070, December 2010.
9. C. Vaquero, A. Ortega and E. Lleida.: Intra-Session Variablity Compensation and a Hypothesis Generation and Selection Strategy for Speaker Segmentation. In: IEEE ICASSP 2011, pp. 4532–4535, Prague, 2011.
10. R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski.: Rapid Speaker Adaptation in Eigenvoice Space. IEEE Transactions on Audio, Speech, and Language Processing, 8(6), 695–707, November 2000.
11. N. Dehak et al.: Front-End Factor Analysis For Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing, 2010.
12. P. Kenny.: Bayesian Speaker Verification with Heavy-Tailed Priors. In: Odyssey Speaker and Language Recognition Workshop, 2010.
13. C. Vaquero, N. Scheffer and S. Kajarekar.: Impact of Prior Channel Information for Speaker Identification. In: ICB 2009, Advances in Biometrics, Lecture Notes in Computer Science, 5558, 443–453. Springer, 2009.
14. C. Vaquero, O. Vinyals and G. Friedland.: A Hybrid Approach to Online Speaker Diarization. In: Interspeech 2010, pp. 2310–2313, Makuhari, Tokyo, 2010.
15. C. Vaquero, A. Ortega and E. Lleida.: Partitioning of Two-Speaker Conversation Datasets. In: Interspeech 2011, pp. 385–388, Florence, 2011.
16. The NIST Year 2008 Speaker Recognition Evaluation Plan, `http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf`
17. C. Vaquero, A. Ortega, A. Miguel and E. Lleida.: Quality Assessment for Speaker Diarization and its Application in Speaker Characterization. Submitted to: IEEE Transactions on Audio, Speech, and Language Processing. Available at: `http://vivolab.es/publications/Qasses4SPKDiar.pdf`.

Perceptual Optimization of Unit-Selection Text-to-Speech Synthesis Systems by means of active interactive Genetic Algorithms

Lluís Formiga* and Francesc Alías

Enginyeria i Arquitectura La Salle,
C/Quatre Camins, 20 - 08022 Barcelona
{llformiga,falias}@salle.URL.edu
http://www.salle.URL.edu

Abstract. The tuning process of Unit Selection TTS (US-TTS) system is usually performed by an expert that typically conducts the task of weighting the cost function by hand. However, hand tuning is costly in terms of the required training time and inaccurate and ambiguous in terms of methodology. With the purpose of easing the task of properly tuning the weights of the cost function, this thesis make its contribution from a perceptual-based approach using of active interactive Genetic Algorithms (aiGAs). The thesis pursues four major guidelines: *i*) accuracy when tuning the weights, *ii*) robustness of the obtained weights, *iii*) real world applicability of the methodology to any cost function design, and *iv*) finding consensus of the different users when tuning the weights. The experimentation is carried out through a small and medium sized corpus (1.9h) applied to different configurations (type of features) of the US-TTS cost function. The thesis concludes that aiGAs are highly competitive in comparison to other weight tuning techniques from the state-of-the-art.

Keywords: speech synthesis,unit selection,weight tuning,interactive evolutionary computation,human computer interaction,latent models

1 Introduction

State-of-the-art speech synthesis strategies base their methodology in the use of large speech corpora with the aim to obtain high-quality synthetic speech. Corpus based TTS are considered third generation of TTS systems. The two main corpus-based strategies are Unit-Selection (US) [1] and Hidden Markov Models (HMM) [2]. US-TTS synthesis is based on the waveform concatenation of acoustic units retrieved from very large speech corpora, while HMM-based TTS (hereafter named HMM-TTS) builds an statistical parametric model of the acoustic units from a smaller amount of speech data.

US-TTS systems are able to achieve very high quality synthetic speech, both in terms of naturalness and expressiveness, when the target message matches the acoustic characteristics of the recorded speech. However, the synthetic speech shows a significant quality decrease when the US-TTS system is asked to generate out-of-domain

* The thesis document might be downloaded at http://www.tdx.cat/handle/10803/21796 and the presentation slides at http://www.slideshare.net/llformiga/llus-formiga-phd-thesis

or out-of-coverage speech (i.e., far from the speaking styles contained in the speech corpus).

The classical expert-based tuning techniques generally finish the weight optimization problem by finding a *unique* set of weights (weight vector) for the whole corpus, without considering that the relative importance of the costs may vary depending on the contextual and phonetic specificity of the acoustic units when being selected. This issue has been previously identified by the literature related to objective weight tuning [3], but as far as we know this issue has not still been addressed by perceptual weight tuning strategies. Hence, developing a weight tuning strategy that respects the units contexts besides embedding subjective preferences could be of great interest.

This thesis studies an human-aided evolutionary strategy of perceptual weight-tuning that combines active interactive genetic algorithms (aiGA) with a clustering step in order to respect the contextual and phonetic features that present a similar behavior. In particular, the thesis first studies a proof-of-concept of the strategy [4], where the aiGA-based proposal was successfully applied to obtain reliable subjectively tuned weights under a classical unit selection scheme. In a deeper study, the seminal aiGA-based weight-tuning scheme is improved by adapting it to obtain context-dependent weights under a real unit selection scenario (i.e., with a larger corpus and a complex cost function).

This extended abstract is organized as follows. Section 2 reviews the background of a evolutionary-based strategy for tuning the weights of the cost function. Section 3 explains the motivations and the goals pursued by the thesis. Section 4 explains the main contributions of the thesis. Section 5 details how the overall strategy was validated with respect to other baseline perceptual tuning strategies and the discussion and final conclusions are detailed in Section 6.

2 Background

The abilities of interactive evolutionary computation are well known when it comes to fuse human and computer efforts with the subjective perception of speech [5]. To reduce user's fatigue in the perceptual tuning, and hence limit the noise at evaluation stage, Llorà et al.[6] defined active interactive Genetic Algorithms (aiGA). This approach allows to discard noisy users or sentences, split the optimization task into a collaborative multi-user task, identify the ambiguity of specific target pairs or obtain a complete rankings of solutions. There are, already, works where aiGA have been used for the perceptual optimization of speech applications (e.g., see [5]).

Figure 1 presents the core execution flow of the proposed aiGA-based method for tuning the cost function weights of US-TTS synthesis systems [4]. Given an initial set of different synthesized versions of the same stimuli (obtained from different weights), these are presented to the user in a tournament hierarchy, then he/she listens to them and selects the winners according his/her subjective criteria. Then, the preferences of the user in front of the proposed synthetic pairs are collected, and a partial-ordering graph is incrementally built by adding this round of user preferences. This graph is used to compute the synthetic fitness function described on the previous section. Once it is available, the continuous-variable based PBIL (cPBIL) optimizes the fitness. The

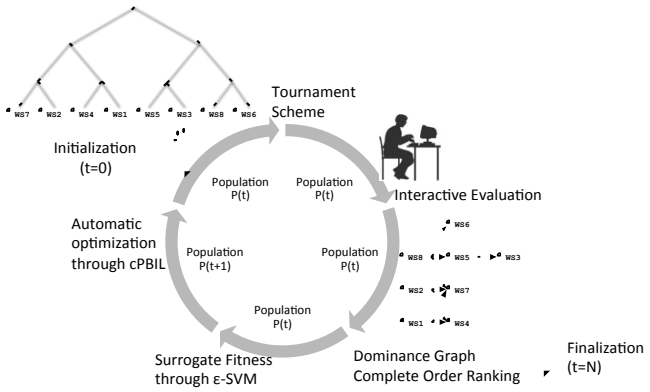


Fig. 1. Process diagram of the aiGA-based weight tuning strategy for US-TTS.

important output of this process is a probability distribution over the weight configurations. This probability distribution models the current user preferences towards good solutions. The new round of solutions to be presented to the user is a fifty-fifty combination of previous shown top ranking solutions, and new solutions sampled out of the learned probability distribution—representing promising solutions according to the observed user preferences, also known as *educated guesses*. The process is repeated for three iterations, as our previous work showed it was a good stop criterion [4].

Following this scheme, the pairwise comparisons that would be necessary to establish a complete ranking of the weighting alternatives (N) for a particular sentence is dramatically reduced with respect to an exhaustive pairwise comparison scheme ($\frac{N \cdot (N-1)}{2}$), e.g., for $N = 16$, the number of required evaluations is reduced by more than 87% ($15 < 120$).

Several approaches have been introduced in acoustics research for combining human global evaluation based on ranking and automatic optimization schemes [7]. However, these works never addressed the problems of human contradictions and ambiguity. On the other hand, other schemes based in graph theory allow to obtain more complex schemes addressing this human interaction problems and therefore reduce the user fatigue by means of *educated guesses* of the user preferences.

3 Motivation and goals: Accuracy, robustness, consensus and real world applicability

The seminal idea of applying an evolutionary strategy for tuning the weights needed a more thorough investigation. Prior to the thesis, the research has bounded to assess the feasibility of the strategy through a small prototype under controlled environment [4]. Concretely, experiments were performed with a extremely small corpus (8 min.) with manual supervision and a single type of unit selection costs (acoustic).

The main objective of the thesis was to design a methodology to perceptual cost function that satisfies the principles of accuracy, robustness, consensus and real world

applicability of US-TTS. In this sense, the thesis presents several contributions to meet each of these principles. Once the new methodology is defined the thesis studies its suitability in a real selection scenario making and comparing it with other state-of-the-art tuning strategies, focusing on the competitiveness, efficiency and significance of the final quality of the synthetic speech obtained.

Next, the four important aspects pursued by the thesis are presented.

3.1 Accuracy

Prior to the thesis, human-aided tuning was only considered towards obtaining a *unique* weight combination regardless the different phonetic or contextual characteristics where the selection of the units takes place [4]. In contrast, the classical methods for tuning the weights that do not have human intervention, are considered at unit level. That is that a specific, and different, weight configuration is applied when recovering the units with the aim to generate a more naturally sounding voice (e.g. the relative importance of the duration near the stops). Hence, the appropriate weighting should be specified by type of the unit to be selected (e.g. occlusive vs. liquid) [3]. However, under a perceptual approach, it is unfeasible to obtain weights specifically for each unit, and therefore, to obtain fine-grained human-supervised knowledge of the relative importance of each feature. Thus, perceptual approaches only obtain weights with low accuracy but qualitatively good in terms of prediction. This issue arises the first motivation of the thesis: set a methodology that allows to obtain high-quality and fine-grained perceptual weights adapted to the different types of contexts and units that take place within the cost function.

3.2 Robustness

Robustness of automatic optimization: Among all the state-of-the-art approaches to perform the weight tuning, the most wide-accepted method is to perform a linear regression between the different cost features and a qualitative estimate by means of Cepstral distances [8]. However regression estimates (RMSE and R^2) yield poor regression models due to the noise and ambiguity of the task. The thesis makes a thorough study of different data preprocessing schemes in order to overcome the noise and obtain more reliable models. In addition, other optimization techniques such as evolutionary strategies are studied as they are well-known for their robustness with respect to noisy scenarios. Up to date of the thesis, we have not found in the literature exhaustive studies that address the issue of cost function feature normalization.

Robustness of human-aided optimization The aiGA prototype [4] highlighted the problem of contradictions of different users yielding to some unreliable models. Therefore, it emerged the need of a methodology for assessing the quality of the solutions provided by the users according to their consistency. Furthermore, the thesis studies other aspects like ambiguity: a user might be consistent, and yet have a passive or heavy hesitant attitude providing excessive evaluation ties compared to other users (evolutionary ambiguity). In addition, the thesis analyzes, the evolution of the user towards a single or multiple criteria and also quantifies the level of consensus among different users.

3.3 Real world applicability:

The seminal prototype [4] demonstrated the feasibility of the aiGA only under a controlled environment (8 min. speech) and poorly populated cost function: only F0, energy, duration and spectral coefficients obtained from the recorded signal. In addition, the annotation of the unit selection corpora was fully supervised avoiding classical errors from an automatically annotated corpus. Thus, the prototype scenario was far beyond from a real selection scheme. The thesis fulfills the need of studying the applicability of the aiGA scheme in a real selection scenario with a larger corpus (at least > 1h) and with a cost function that considers different type of features (acoustic, linguistic, contextual...).

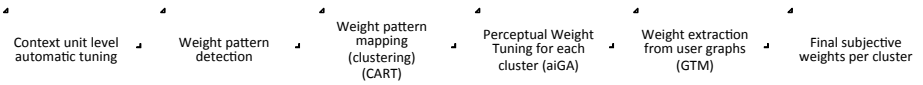


Fig. 2. Step diagram of the proposed methodology to tune weights at context-dependent cluster level.

3.4 Consensus

When different users evolve the same problem, the integration of the different solutions arises a problem when it exist some contradiction in the criteria (eg. intelligibility vs. similarity to a natural voice). Alías et al. [4] solved this problem by a second pass by the users, where the best solutions obtained by aiGA were competitively selected by other users. In this sense, it was pointed the need of a method able to integrate the different criteria of the users. This issue was also identified by the creators of aiGA [9]. Whereas individual users may disagree about the best solution to the problem, the multiple criteria can be mapped into a single global model, and therefore obtain a consensus model that satisfies the preferences of different users. In this work we study this problem using models with latent variable, which solve the problem of contradictions under a Gaussian non-heuristic adaptive approach.

4 Main contributions: Adaptive Weight Tuning through active interactive Genetic Algorithms

4.1 Perceptual weight tuning at context-dependent cluster level

The first contribution of thesis (namely aiGAClustered) allows to obtain perceptual weights and respect the phonological, linguistic and contextual specificity of the selected units.

Generally, weight tuning can be addressed at three levels of precision (see figure 3) [1]: at unit level [10, 8] (one weight pattern for each phoneme), at cluster level [4] (one weight pattern per group of units, e.g., fricatives), or at global level [11] (a unique

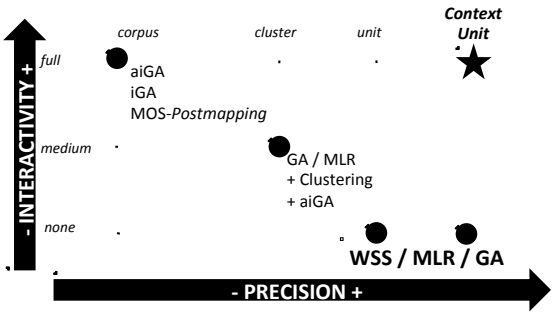


Fig. 3. Context-dependent reformulated weight-tuning levels depending on the interactivity and precision offered by the tuning strategy (the star shows the desired setting).

weight pattern for *all* the units contained in the corpus). Generally, the level of adjustment is directly related to the technical difficulty of the tuning process: weights are typically tuned at unit level when the process is automatic and without human intervention (e.g., MLR or GA), the tuning is performed at global level when it is performed perceptually through a reduced set of representative utterances, both for computer-aided [4, 11] and expert-based hand-tuning approaches [12].

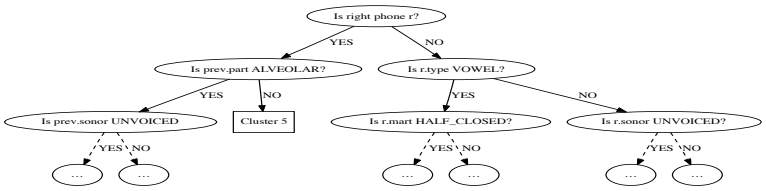


Fig. 4. Question set CART-clustering in order to determine weight groups.

Therefore, the thesis considers that similar units can be grouped according to their common behavior (pattern) when tuning weights automatically. Afterwards the weights may be refined through a perceptual tuning stage (see figure 4). It is important to highlight that clusters are only obtained to detect weight behavior patterns, but they do not decide the final weight values of the group as other approaches do [13]. The strategy obtains an intermediate level of accuracy (see figure 3 between global (all units together) and unit level tuning, respecting the inherent specificity of each unit [3]. Hence, weight tuning at cluster level overcomes the drawbacks of fatigue and design complexity that are typically involved in perceptual tuning [4].

The proposed methodology is as follows: *i*) initial weights are obtained by means of an automatic methodology (e.g. GA) – that is a different weight value for each unit in a specific context. *ii*) contextual units with similar weight values are grouped (see figure 4) with a state-of-the-art algorithm (we have chosen Expectation Maximization after a detailed comparative study among other techniques). *iii*) a classification tree (CART) algorithm [10]) is used to map the clusters to the distinct contextual scenarios of the

cost function. The CART question set was defined to include phonological (e.g. point of articulation),linguistic (e.g. POS) and contextual information for each half-phone of the corpus. The number of clusters was been constrained by analyzing several impurity measures and selecting the number of clusters that obtained the most favorable validity scores [4].

4.2 Generative Topographical Maps to find consensus models

The second contribution of this thesis focuses on the study of integration (consensus) of the different solutions evolved by the users. Previously [4], due to the lack of a methodology to automatically obtain consensus, it was necessary a second pass to choose the best weights among the different solutions. The thesis studies the combination of user preferences at genotype level, i.e. in terms of weight vectors. Although some proposals of aiGA consensus have been proposed at phenotype (graph combination) level [9], it did not solve the problem of breaking the cycles (inconsistencies) formed by the different users, at least without heuristics. The consensus by genotype allows to search for the best solution with a Gaussian adaptive approach that fits to the data (weight values) distribution.

In this sense, this thesis proposes latent models in order to model and group the best preferences of evolving users and therefore extract the best set of weights from the model. The rationale after the latent models is motivated by their capacity to find models beyond linear dependencies of the data and their EM strategy makes them robust to noise. Furthermore, Gaussian latent methods implicitly deal with the contradictions between different users without having to resolve them heuristically. After some experimentation, the thesis concludes that generative topographic maps (GTM) are able to integrate different preferences robustly (thanks to the EM adaptation). That is coherent with previous work carried out in speech processing, as it confirms that Gaussian-EM based algorithms are good to deal with speech problems either for *i*) detecting patterns of weights (with EM) and em ii) finding consensus models, as GTM is a constrained version of GMM.

4.3 Evolutionary Process Indicators

Prior to the thesis [4] there was only a single indicator of consistency defined (κ) which measured the degree of consistency of user preferences graph model. During the experimentation of the thesis, we observed this measure as insufficient to obtain all the necessary information about the quality of the solutions. Thus, we expanded the aiGA indicators to: three new indicators: *i*) A measure that about the confusion within the graph (Certainty Ratio λ), *ii*) A measure about the convergence of the perceptual test to a single or multiple solutions (*Intra-user* Convergence Ratio ρ) and *iii*) A measure about the consensus of the test conducted by different users (*Inter-user* Correlation Ratio τ). The new indicators allowed to discard runs that had not converged or didn't correlate with other users' runs and detect whether the graph building process had to be stopped prematurely because it was not evolving.

4.4 Other contributions

Other remarkable contributions of the thesis that were a byproduct of the performed experimentation are: *i*) a detailed study of different preprocessing, normalization and transformation strategies to obtain more reliable models. *ii*) the demonstration of the capabilities of *aiGAClustered* methodology under a real-world US-TTS scenario with a large training corpus with automatic annotations prone to errors and *iii*) the suitability of the methodology to complex cost functions (with a large amount and different type of cost features). Rigorous experimentation and analysis has been carried out to all the contributions explained.

5 Methodology Validation

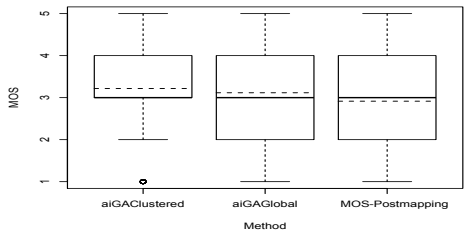


Fig. 5. MOS Comparison of different methods of perceptual weight tuning studied.

The weights obtained by the methodology proposed in the thesis (aiGA at context-dependent level) were compared to a perceptual baseline approach (MOS-Postmapping – [11, 11]). MOS-Postmapping tunes a single weight combination for all the different units and contexts in the corpus (global granularity).

Furthermore, we also extracted a single global weight combination from aiGA graphs,i.e. for all contexts and units. This evaluation scheme allowed completing the study in two aspects: *i*) the impact of the clustering strategy to the perceptual tuning methodology (Global aiGA vs. Clustered aiGA), and *ii*) a fair comparison between aiGA and MOS-Postmapping working both at the same global level of tuning.

The validation of the three strategies was conducted with 33 users (including the 8 expert tuning evaluators), considering different levels of speech-related expertise between users throughout 10 randomly sentences and 3 weight configurations (Global aiGA, Clustered aiGA and MOS-Postmapping). That made a total of 30 pairwise comparisons for each user, collecting a total of 990 pairwise annotations and 1980 absolute MOS scores.

Figure 5 depicts the results obtained on the MOS scale considering the preferences of the users. The absolute scores obtained were as follows: $MOS_{aiGAClustered} = 3.21$, $MOS_{aiGAGlobal} = 3.11$ and $MOS_{MOS-Postmapping} = 2.91$. The direct comparison of the different configurations (via a CMOS-like sign analysis of the pair-wised absolute scores) revealed that all differences were significant

In absolute terms, results showed a clear preference for the clustered aiGA methodology as is preferred on 96/241 pairs (39.83%). That number of votes was significantly better than the 74/241 votes (30.70%) obtained by the global aiGA and the 71/241 votes (29.46%) by the MOS-Postmapping method. ($p < 5 \cdot 10^{-}3$).

6 Discussion and Final conclusions

The thesis describes context-dependent methodology based on active interactive genetic algorithms for the perceptual tuning of the cost function weights of TTS systems including unit selection (both US-TTS and hybrid TTS systems). The seminal aiGA-based methodology was improved from the original proof-of-principle scheme (small corpus and simple cost function) to a real unit selection scenario (large corpus and complex cost function). Therefore, the aiGA-based strategy is a robust strategy when combining different kind of weights (linguistic vs.ãcoustic) for different types of costs (discrete vs.c̃ontinuous). In addition the thesis makes two contributions that are crucial to find weights that improve significantly the quality of the synthesized speech: *i*)Evolutionary process indicators of aiGA and *ii*) the application of a consensus modeling method (GTM). Moreover, the experiments demonstrate the importance of considering the context of units when it comes to tune the weights. *aiGAClustered* was preferred in front of other strategies in the perceptual preference tests involving different ways of user-interaction and tuning level grain. Therefore, the context-dependent aiGA-based is an excellent tool for tuning different weights perceptually respecting the contextual and phonetic specificity of the involved units.

The research developed under the framework of this thesis has had a considerable impact in the community involving: 1 peer-reviewed journal [4], 5 conferences related to Speech Processing [14–18], 2 conferences related to Artificial Intelligence [19, 20] and one book chapter [21].

References

1. Hunt, A., Black, A.W.: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Volume 1., Atlanta (USA) (1996) 373–376
2. Tokuda, K., Zen, H., Black, A.: An HMM-Based Speech Synthesis System Applied to English. In: Proc. of IEEE Workshop on Speech Synthesis, IEEE (2003) 227–230
3. Campillo, F., Alba, J.L., Rodríguez-Banga, E.: A Neural Network Approach for the Design of the Target Cost Function in Unit-Selection Speech Synthesis. In: Proc. of the 9th International Conference on Speech Communication and Technology (InterSpeech), Lisbon (Portugal), ISCA (2005) 2533–2536
4. Alías, F., Formiga, L., Llorá, X.: Efficient and Reliable Perceptual Weight Tuning for Unit-Selection Text-to-Speech Synthesis based on active interactive Genetic Algorithms: A Proof-of-Concept. Speech Communication **53**(5) (2011) 786–800
5. Alm, C., Llorà, X.: Evolving Emotional Prosody. In: Proc. of 9th International Conference on Spoken Language Processing (ICSLP), Pittsburgh, PA (USA), ISCA (2006) 1826–1829
6. Llorà, X., Sastry, K., Goldberg, D.E., Gupta, A., Lakshmi, L.: Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. Proc. of the Genetic and Evolutionary Computation Conference (GECCO) (July 2005) 1363–1371

7. Lee, M., Lopresti, D., Olive, J.: A Text-to-Speech Platform for Variable Length Optimal Unit Searching using Perception Based Cost Functions. International Journal of Speech Technology **6**(4) (2003) 347–356
8. Meron, Y., Hirose, K.: Efficient Weight Training for Selection based Synthesis. In: Proc. of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Volume 5., Budapest (Hungary) (1999) 2319–2322
9. Llorà, X., Yasui, N.I., Goldberg, D.: Graph-theoretic measure for active igas: Interaction sizing and parallel evaluation ensemble. In: Proc. of the 10th Conference on Genetic and Evolutionary Computation (GECCO), Nova York (USA), ACM (2008) 985–992
10. Black, A.W., Taylor, P.: Automatically Clustering Similar Units for Unit Selection in Speech Synthesis. In: Proc. of the 5th European Conference on Speech Communication and Technology (EuroSpeech), Rhodes (Greece) (1997) 601–604
11. Chu, M., Peng, H.: An Objective Measure for Estimating MOS of Synthesized Speech. In: Proc. of the 7th European Conference on Speech Communication and Technology (EuroSpeech), Aalborg (Denmark), ISCA (2001) 2087–2090
12. Clark, R., Richmond, K., King, S.: Multisyn: Open-Domain Unit Selection for the Festival Speech Synthesis System. Speech Communication **49**(4) (2007) 317–330
13. Colotte, V., Beaufort, R.: Linguistic Features Weighting for a Text-to-Speech System Without Prosody Model. In: Proc. of the 9th European Conference on Speech Communication and Technology (EuroSpeech), Lisbon (Portugal), ISCA (2005) 2549–2552.
14. Alías, F., Llorà, X., Iriondo, I., Formiga, L.: Ajuste subjetivo de pesos para selección de unidades a través de algoritmos genéticos interactivos. Procesamiento del Lenguaje Natural **31** (September 2003) 75–82
15. Alías, F., Llorà, X., Iriondo, I., Sevillano, X., Formiga, L., Socoró, J.: Perception-Guided and Phonetic Clustering Weight Tuning Based on Diphone Pairs for Unit Selection TTS. In: Proc. of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju Island (South Korea), ISCA (2004) 1221–1224
16. Alías, F., Llorà, X., Formiga, L., Sastry, K., Goldberg, D.E.: Efficient Interactive Weight Tuning for TTS Synthesis: Reducing User Fatigue by Improving User Consistency. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Volume I., Toulouse (France), IEEE (2006) 865–868
17. Monzo, C., Formiga, L., Adell, J., Iriondo, I., Alías, F., Socoró, J.: Adaptación del cth-url para la competición albayzin 2008. In: Actas de las V Jornadas en Tecnologias del Habla. (2008) 87–90
18. Formiga, L., Trilla, A., Alías, F., Iriondo, I., Socoró, J.C.: Adaptation of the URL-TTS system to the 2010 Albayzin Evaluation Campaign. In: Proc. of Fala 2010, VI Jornadas en Tecnología del Habla, Vigo (Spain) (2010) 363–370
19. Formiga, L., Alías, F.: Extracting User Preferences by GTM for aiGA Weight Tuning in Unit Selection Text-to-Speech Synthesis. Lecture Notes in Computer Science - Computational and Ambiental Intelligence **4507** (June 2007) 654–661 Proc. of the 9th International Work-Conference on Artificial Neural Networks (IWANN).
20. Formiga, L., Alías, F., Llorà, X.: Evolutionary Process Indicators for active iGAs Applied to Weight Tuning in Unit Selection TTS Synthesis. In: Proc. of the IEEE Conference on Evolutionary Computation (CEC), Barcelona (Espanya), IEEE (July 2010) 2322–2329
21. Formiga, L., Alías, F.: GTM User Modelling for aiGA Weight Tuning in TTS Synthesis. In: Encyclopedia of Artificial Intelligence. Information Science Reference (2008) 788–795

IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

Oral Session O5:
Robustness in Automatic Speech Recognition

IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

Pairwise Likelihood Normalization-Based Channel Selection for Multi-Microphone ASR

Martin Wolf and Climent Nadeu

TALP Research Center,
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
{martin.wolf,climent.nadeu}@upc.edu

Abstract. In scenarios where several distant distributed microphones are capturing signal in parallel, the quality of the recorded speech is different for each channel. Hence, some signals may be more suitable for further processing by an ASR system than others. To benefit from this fact, a good indicator of signal quality must be designed. In this paper, after explaining the problems associated with the use of likelihood as quality measure, a channel (microphone) selection technique based on pairwise likelihood normalization is introduced. When compared in digit ASR experiments with other channel selection techniques published so far, it shows comparable or better performance, provided that there is enough speech data to reliably estimate the quality measure.

Keywords: Automatic speech recognition, reverberation, channel selection, acoustic likelihood, pairwise normalization

1 Introduction

Besides background noise, reverberation strongly affects the performance of Automatic Speech Recognition (ASR) systems when using distant microphones in enclosed environments. Acoustic waves reflected by the walls and the objects in the room arrive to the microphone attenuated and with different delays, so introducing undesirable interferences. In scenarios where several distant distributed microphones are capturing signal in parallel channels, the distortion of the acquired speech varies. Consequently, depending on which channel is used for the recognition, the performance of ASR system may also vary. Channel Selection (CS) takes advantage of this fact and tries to select the channel which leads to the lowest Word Error Rate (WER) after recognition.

The selection decision may be done in two ways. First, the least distorted signal may be selected for recognition by assuming that the ASR accuracy depends on the quality of the captured speech signal. In such case we talk about signal-based CS methods. The signal (channel) distortion can be estimated using some signal-processing measure. In [1] and [2] the Signal to Noise Ratio (SNR) was used for this purpose. In [3], the possibility to use information about the relative position and orientation of speaker and microphone for CS was discussed. The method is based on the assumption that speech should be less distorted

by reverberation if the microphone is closer to the speaker. In [4], the relations between different parts of the Room Impulse Response (RIR) and the resulting WER of the ASR system were investigated. It was shown that certain parts of the RIR harm the speech recognition more than others. Based on that, we proposed in [5] a CS method that uses a measure extracted from the channel characteristics, which is calculated from the RIR as the energy of the late reflections (those harming ASR the most) normalized by the energy of the whole RIR. This technique assumes that the RIRs can be accurately estimated.

To avoid the problem of RIR estimation, we proposed another CS method, where the distortion measure is extracted directly from the speech signal. It is a well known fact that reverberation smoothes the energy time envelope [6], so the effect of reverberation may be observed as a reduction in the variance of that envelope. We measure this reduction calculating the variance of the speech energy envelope in the sub-bands [3].

The second group of CS techniques are so called decoder-based methods. In them, the decision measure is either directly related to the decoding part of the recognition system (e.g. likelihood, posterior probability) or is obtained from some type of classification. A straight forward decoder-based approach was presented in [7]. Signals from all channels are passed to the recognizer and the channel giving the maximum likelihood is selected. In [1], some feature compensation technique (e.g. mean and variance normalization or histogram equalization) is first applied to each channel. Then, both the original and the compensated feature streams are recognized, and the channel with the smallest difference between the recognized word sequences from the original and the compensated versions is selected. In [8], CS based on a class separability measure was introduced. The channel maximizing that separation measure from speech feature vectors is selected for recognition.

CS may be useful whenever there are two or more signal streams and we have to select one for recognition. The ASR system may further benefit from CS for two reasons. Firstly, the spatial diversity of the room is exploited and the least distorted signal is chosen, so it may be less difficult to clean (dereverberate) it or to decode the correct sequence. Secondly, mainly the signal-based CS methods may be used to reduce the number of channels and so decrease the computational load of subsequent elements in the recognition chain as was for example done in combination with beamforming in [9].

In this paper we focus on the decoder-based approaches that use measures related to the acoustic likelihood. First, we discuss in Section 2 the problem associated to the use of a non-normalized likelihood, and propose a likelihood-ratio based measure (pairwise normalization) to circumvent the problem of computing posterior probabilities. In Section 3 we evaluate it experimentally on a digit recognition task, and compare it with the decoder-based CS methods presented in [1] and [7], and finally with the signal-based method from [3].

2 Likelihood-based CS

The first, straight forward, decoder-based approach based on acoustic likelihood was presented in [7]. In that implementation, signals from all channels were passed to the recognizer and the channel giving the maximum acoustic likelihood was selected. In the following, we will outline a problem related to this measure and explain why it should not be used for CS. A solution is provided in the next subsection.

In conventional ASR systems the Bayes' rule is used to compute the posterior probability $P(w \mid O)$ of a word sequence $w = w_1, \ldots, w_N$, given a sequence of acoustic observation vectors $O = o_1, \cdots, o_T$. The speech recognition problem is therefore regarded as that of computing

$$\begin{aligned} W &= \arg\max_{w \in \Omega} P(w \mid O) \\ &= \arg\max_{w \in \Omega} \frac{p(O \mid w) P(w)}{p(O)}, \end{aligned} \tag{1}$$

where Ω denotes the set of all possible word sequences, $p(O \mid w)$ is the acoustic likelihood, and $P(w)$ is the prior probability of the word sequence. The probability of the observation vector $p(O)$ in the denominator is usually omitted since it does not depend on the selected sequence of words and works only as a scaling factor. Thus, the use of (non-normalized) likelihood scores is not a problem in the case of a single channel.

If there are several parallel channels, the probability of the observation vector $p(O_m)$ is different for each stream. The posterior probability for the multi-channel case is defined as

$$P(w \mid O_m) = \frac{p(O_m \mid w) P(w)}{p(O_m)}, \tag{2}$$

where $m = 1, \cdots, M$ is the channel index. Obviously, the probability of the observation vector can not be neglected if we want to compare the scores from different channels among themselves. This is the reason why likelihoods provided by ASR systems can not be used as a reliable indicator of channel quality.

As mentioned at the beginning of this subsection, likelihood was applied to the channel selection problem in [7]. In that work, a set of acoustic models was trained for each speaker-to-microphone distances. In the first step, a maximum likelihood criterion was used to select the best matching set of models for the observation sequence in each channel. Non-normalized likelihoods may be used for that, because they are computed using the observations from the same channel. In the second step, the likelihood scores from the channels were compared, and the channel with the maximum likelihood was selected. If likelihoods are not normalized, according to the reasoning above, the comparison is not meaningful. Authors do not mention any explicit normalization, neither discuss this issue, so we can only assume that non-normalized values were used. In our opinion, the accuracy improvements reported in [7] may be attributed to the selection

of the proper set of models (i.e. the selection of the speaker-to-microphone distance) that is carried out in the first step. Moreover, we can see in the reported results that after that selection the performance is already quite high in almost all channels. So even if the channel was selected randomly in the second step, the degradation would not be strong. In the following we will discuss a possible solution to the problem outlined here.

2.1 Pairwise likelihood normalization

If the posterior probability (2) was available for each channel, it could be used to select the most likely word sequence among all channels. However, the estimation of the observation vector probability $p(O_m)$ would be required for each channel, an objective that may be unattainable since the number of possible channel transfer functions may be very large, or even unlimited when the speaker is moving. In the following we will present an alternative way that is based on the ratio of likelihoods.

Let's first discuss the case with 2 channels. For simplification purposes, we will assume equal prior probabilities for all word sequences and neglect them in (2). Assume we know the decoded sequences w_1, w_2 and their corresponding likelihoods $p(O_1 \mid w_1)$, $p(O_2 \mid w_2)$ for both channels. If the word sequences are the same, the channel is selected randomly and no further CS related processing is needed. If they are different, we may take the sequence w_1 from the first channel and compute the likelihood of the observation from the second channel $p(O_2 \mid w_1)$ and vice versa, $p(O_1 \mid w_2)$. Then, by dividing the likelihood $p(O_1 \mid w_1)$ by the likelihood corresponding to the alternative sequence $p(O_1 \mid w_2)$, we get a measure of confidence about the correctness of the decoded sequence w_1 in that channel 1. After computing the corresponding likelihood ratio for channel 2, the channel with maximum ratio (confidence) is selected.

For the general multi-channel case, by using addition of likelihood ratios across channels, the selection criterion can be expressed as

$$C = \arg\max_m \sum_i \frac{p(O_m \mid w_m)}{p(O_m \mid w_i)}, \tag{3}$$

where $m = 1, \cdots, M$ and $i = 1, \cdots, M$ are channel indexes, and C is the selected channel. By doing this, we are selecting the channel which shows in average the maximum confidence regarding to its decoded word sequence. A computational drawback of this method is that in case all the M decoded sequences are different, the likelihood of the alternative hypothesis has to be calculated $M - 1$ times.

3 Experiments

3.1 Experimental setup

The experiments were conducted using TIDIGITS, a well known database of connected digits in English. For testing, the original close-talk recordings were

convolved with real RIRs from two different rooms. For the first set of RIRs [5], the measurements were made in the UPC Smart-room shown in Fig. 1, using a sweep excitation signal with a logarithmically increasing frequency. There are 6 T-shaped omni-directional microphone clusters with 4 microphones per cluster installed on the walls. In our experiments we selected only the microphone in the middle of each cluster. Seven different positions were chosen in the room, as indicated in Fig. 1, and four orientations of the loudspeaker were defined for each position, simulating the scenario where a person is giving a talk and moves along the room. The Reverberation Time (RT) T_{60} of the room is around 530ms.

The second RIR set [10] was recorded in a room called Octagon. The RT is much longer than in the previous case (frequency dependent RT T_{30} for 125 Hz goes up to 4s), and the position and orientation of the sound source is unique. A microphone was placed in several positions in a grid-like fashion, as shown in Fig. 2, covering most of the octagonal room space. It is quite evident in this case which channels should be preferred in terms of distortion, since the recordings with the microphone placed right in front of the speaker will probably be less harmed by reverberation. For the experiments, we use only a subset of the microphones that are marked with the full black dot in Fig. 2.

The sampling frequency in the experiments was 16 kHz. We used continuous density hidden Markov models (HTK toolkit) based system, with 3 Gaussians per state for the words, and 6 for the silence model. The 11 models for words (digits zero, oh, one, ..., nine) have 16 states, there are 3 states for the silence model, and one for a short pause model which is shared with the middle state of the silence model. Models were trained only with clean speech. Standard Mel-Frequency Cepstral Coefficients (MFCC) calculated from 20 mel-frequency bands were used. The feature vector included 12 cepstral coefficients without the 0^{th} coefficient, frame energy, delta and acceleration features, so size of the vector was 39. Frame length was 25ms and frame shift 10ms. Finally, Mean and Variance Normalization (MVN) was applied.

The frequency by which a new channel is selected is an important parameter. Ideally, the CS algorithm should be able to track the changes in the room and select the channel any time conditions change. If the changes are slow or not very significant we can wait, collect more speech data, and make the measurement more robust. But, if we wait too much, the recognition delays. Also, if the speaker moves, the conditions in the room change so the selection decision may not be optimal. Ideally we would like to have a technique that requires small amount of data to make an estimation of the quality measure and still leads to a good recognition performance.

We designed an experiment, where we vary the frequency by which the channel is selected. First we evaluate the recognition performance selecting the channel per single utterance and then we continuously increase the number of utterances. There are 8700 files in the testing set. We created the groups of utterances of equal size (factors of 8700). This means that in the first test we had 8700 single utterances, next there were 4350 groups with 2 utterances, 2900 groups with 3 utterances, etc. In each experiment the CS measure was calculated from all files

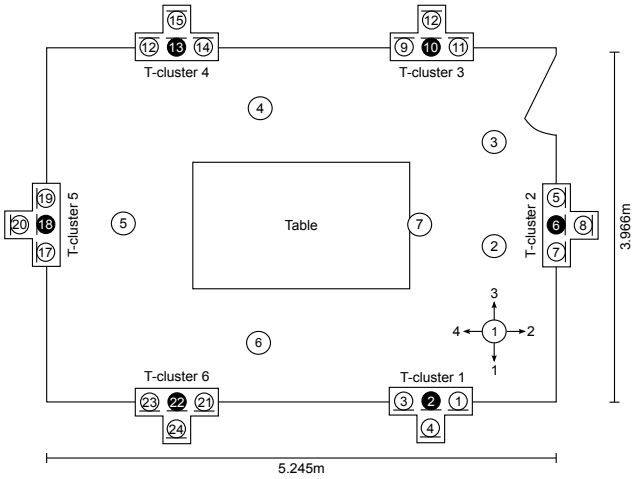


Fig. 1. The UPC Smart room with 24 microphones grouped in T-shape clusters. There are 7 acoustic source positions and 4 orientations.

in a given group, so using progressively more and more data. The same channel was selected for all utterances in that group.

3.2 Evaluated CS methods

In the experiments we compare the likelihood-based CS techniques discussed in Section 2 with the alternative CS methods presented in the literature before. The implementation details are provided here.

To extract the likelihood measure, each utterance was decoded individually and the likelihood for a group or utterances was calculated as the sum of acoustic log likelihoods of individual utterances. Before summation, the likelihoods were normalized by the number of frames to compensate for the different utterance lengths. The channel with the highest accumulated score was selected for that group. The normalized likelihood was extracted in the same way, but the pairwise normalization (see Subsection 2.1) was applied before the summation.

The likelihood-based techniques are further compared with two alternative CS methods. The first one is the decoder-based method that is using Feature Normalization (FN) and was introduced in [1]. In this method a FN technique is applied for each channel and the speech recognition hypothesis before and after FN are compared. The underlying assumption is that the higher the distance between the two hypothesized word sequences, more distorted is the signal. We used MVN as feature normalization method in our tests, that was applied to the baseline features (MFCC + energy $+\Delta + \Delta\Delta$). The normalized feature streams were used to extract the WERs displayed in the results, since they offer a better ASR accuracy than the non-normalized ones. If the minimum distance between two compared hypothesized word sequences was the same in two or more channels, the channel was selected randomly. The measure for a group of utterances

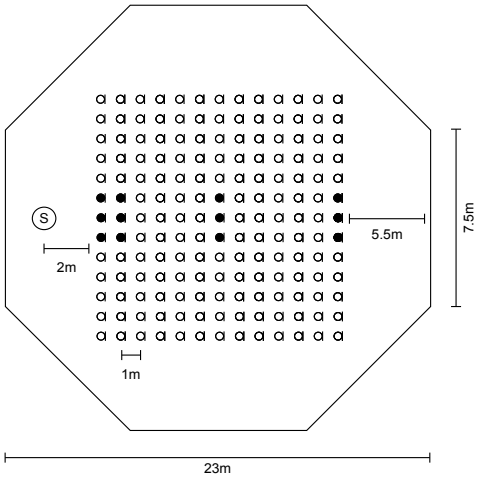


Fig. 2. Speaker and microphone positions in the Octagon room [10].

was in this case calculated as the sum of distances between the decoded sequences from the original and the normalized feature streams, accumulated over all utterances of the group.

The second CS technique used in the comparison is signal-based. It is using the Envelope Variance (EV) measure and was introduced in [3]. Slight modification of the original implementation was made here. To remove the short term effects (e.g. different electrical gain and impulse response of the microphone) from the signal in each channel, the mean value was subtracted in the log domain from each sub-band. This CS technique is further referenced as EV method. In case group of files was used to extract the measure, reverberated utterances were first joined on a signal level and EV measure was then extracted from these long files.

As a reference, we further include a special case of CS, where knowledge of the correct word sequence is assumed. In this oracle CS method, the channel that shows the lowest average WER over all the testing database is chosen. This method will be further referenced as AVG.

3.3 Recognition results

The recognition results in terms of WER for each of the rooms are presented in Fig. 3 and Fig. 4. On the x-axis, there is the number of utterances in the group in log scale. Horizontal dashed line shows the WER when the best channel in average is selected for all utterances using oracle knowledge of the correct sequence, i.e. the AVG case. Note this line can not be seen in case of the Octagon, because it overlaps with the one corresponding to the EV based method.

We can observe that, for the UPC Smart room, the ASR performance using the CS method based on non-normalized likelihood does not improve, in spite

Table 1. The CS performance in terms of WER for the UPC Smart Room and the Octagon, selecting the channel per utterance.

Scenario	RND	L	NL	FN	EV	AVG
UPC	22.7	20.3 (10.6%)	19.5 (14.1%)	19.3 (15%)	19.2 (15.4%)	17.4 (23.3%)
Octagon	56.1	39.7 (29.2%)	35.8 (36.2%)	41.7 (25.7%)	28.4 (49.4%)	28.5 (49.2%)

that more data are used to estimate the measure. It actually gets slightly worse. In the Octagon case, eventually, the performance reaches the AVG case for the large groups. This may be attributed to the fact that MVN tends to equalize the observation probabilities across channels and therefore the acoustic likelihoods can be compared among themselves (see the reasoning in Section 2).

The pairwise normalization clearly helps. A consistent WER decrease may be observed when the size of the data set for the measure computation increases. Actually, in the UPC Smart room it is the only method that reaches the performance of the oracle AVG case for the larger group sizes. Note also that curves are smoother than those from the non-normalized likelihood. In case of the CS method using FN a consistent decrease in WER can be observed for both rooms. This is also true in case of the EV based method in the UPC Smart room. In case of the Octagon, increasing the group size does not bring any improvement since the EV method works well for any number of utterances.

The first point in the graph, which corresponds to the case when only a single utterance is used to select the channel, is interesting from the practical point of view. Recall, that ideally we would like to have a CS technique that requires small amount of data to make a reliable quality measure estimation. In the command based systems often only one short utterance may be used to do that. In Table 1 the recognition results for this case are presented in more detail. The performance in terms of WER using the CS methods based on non-normalized Likelihood (L), Normalized Likelihood (NL), Feature Normalization (FN) and Envelope Variance (EV) is presented for both scenarios, the UPC Smart room and the Octagon. As a reference we also included a case when the channel is selected randomly (RND). The relative improvement for each CS method with respect to the random case is shown in parenthesis. We may see that EV method, despite its computational simplicity, leads to the best recognition performance in both rooms, followed by the FN in the UPC Smart room and by the NL in case of the Octagon.

4 Conclusions

In this paper we presented CS methods based on the quality measures extracted from the back-end of the recognition system. The problem related to the use of the non-normalized likelihood as selection measure was discussed and a pairwise normalization across different channels that results in an average likelihood ratio was proposed. When compared in our digit ASR experiments with other channel

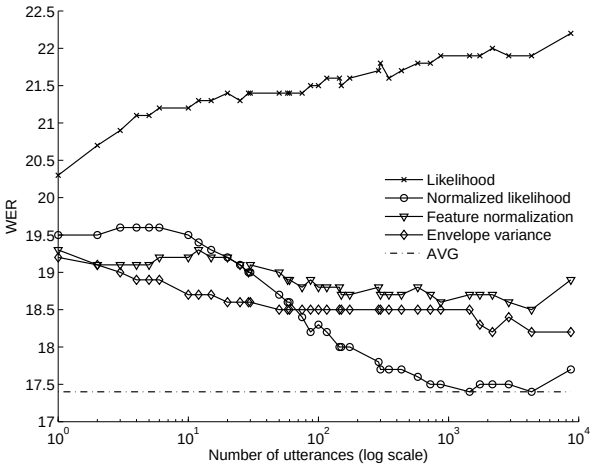


Fig. 3. CS performance. The channel is selected per group of files for the UPC Smart Room

selection techniques, it showed comparable or better performance, provided that there is enough speech data to reliably estimate the quality measure. Therefore, it may be well applied in scenarios where conditions in the room do not change frequently. We further observed that a previously-introduced simple signal-based CS method based on the envelope variance performs better in reverberant conditions than any other of the evaluated techniques in cases, when only small amount of data is available to select the channel.

Decoder-based methods could also be employed in those scenarios which require short recognition delays, if integrated with some measure estimation tracking mechanism. In an initial phase we could accumulate sufficient amount of speech data to get a reliable estimate of the quality measure and then, in the working phase, this estimate could be updated with each new utterance to track the changes in the room. This way we could avoid the recognition latency problem in the working phase, since the CS decision would be updated with each new utterance. The changes in the room conditions could be tracked up to some extend, and we would still keep the advantage of the decoder-based techniques which can work with any kind of distortion.

Acknowledgments. This work was supported by the Spanish project SARAI (reference number TEC2010-21040-C02-01) and the FPU stipend program provided by the Spanish government.

References

1. Obuchi, Y.: Multiple-microphone robust speech recognition using decoder-based channel selection. In: Workshop on Statisitical and Perceptual Audio Processing,

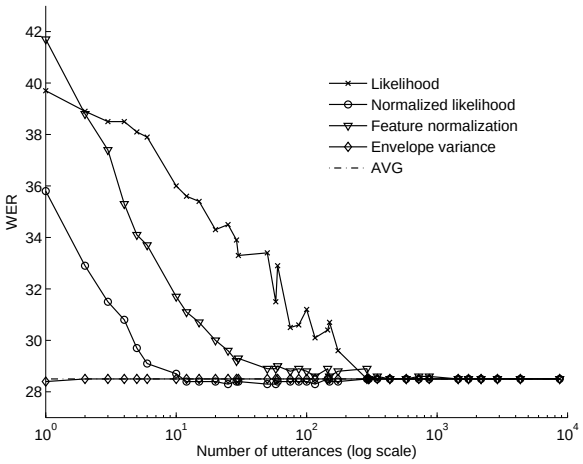


Fig. 4. CS performance. The channel is selected per group of files for the Octagon

Jeju, Korea (2004)
2. Wölfel, M., Fgen, C., Ikbal, S., Mcdonough, J.W.: Multi-source far-distance microphone selection and combination for automatic transcription of lectures. In: Proc. of INTERSPEECH. (2006)
3. Wolf, M., Nadeu, C.: On the potential of channel selection for recognition of reverberated speech with multiple microphones. In: Proc. of INTERSPEECH. (2010) 80–83
4. Petrick, R., Lohde, K., Wolff, M., Hoffmann, R.: The harming part of room acoustics in automatic speech recognition. In: Proc. of INTERSPEECH. (2007) 1094–1097
5. Wolf, M., Nadeu, C.: Towards microphone selection based on room impulse response energy-related measures. In: Proc. of I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal, (2009) 61–64
6. Houtgast, T., Steeneken, H.J.M.: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. The Journal of the Acoustical Society of America **77**(3) (1985) 1069–1077
7. Shimizu, Y., Kajita, S., Takeda, K., Itakura, F.: Speech recognition based on space diversity using distributed multi-microphone. In: Proc. of ICASSP. Volume 3. (2000) 1747–1750
8. Wölfel, M.: Channel selection by class separability measures for automatic transcriptions on distant microphones. In: Proc. of INTERSPEECH. (2007) 582–585
9. Kenichi Kumatani, John McDonough, Jill Lehman, Bhiksha Raj: Channel selection based on multichannel cross-correlation coefficients for distant speech recognition. In: Proc. of HSCMA, Edinburgh, United Kingdom (June 2011) 1–6
10. Stewart, R., Sandler, M.: Database of omnidirectional and b-format room impulse responses. In: Proc. of ICASSP. (2010) 165–168

IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

Albayzin Language Recognition Evaluation
System Descriptions

IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

ANHYT System Description for Albayzin 2012 Language Recognition

Fusheng Ba, Xuefei Jia, Zhihua Jin,
Rui Xing, and Xu Ji

ANHYT, Co., Ltd.,
5th floor, No.9 East Fengde Road,
Yongfeng Industrial Base, Haidian District, Beijing, China
{bafusheng,jiaxuefei,jinzhihua,
xingrui,jixu}@anhyt.com
http://www.anhyt.com

1 Introduction

This paper describes ANHYT systems for Albayzin 2012 LRE. ANHYT submitted 3 systems for 4 conditions. For different conditions, the only difference is the training and development data, which will be clearly described in the following section. The primary system was a fusion of 2 acoustic systems, ivec-LDA and ivec-SVM [1]. The contrast 1 system was the raw score of ivec-LDA. The contrast 2 system was the raw score of ivec-SVM. All of our system configurations strictly followed the Albayzin 2012 LRE's evaluation plan.

2 Primary system

2.1 Train and development data

The Albayzin 2012 LRE released a database named Kalaka-3. This database included three parts: training data, development data and evaluation data.

- For the *Plenty of training*, the primary system took the training data to train the UBM, total variability matrix T, linear discriminant analysis (LDA) matrix and language model. The development data was used for score fusion and calibration.
- For the *Empty of Training*, the primary system used the same UBM and T matrix, but pooled the training data and development data together to train the LDA matrix and language model. The development data was also used for score fusion and calibration.
- For the *Closed-set*, we removed OOS languages to build our systems.
- For the *Empty-set*, we took OOS languages as a mixed language to build our systems.

2.2 Front-end processing

The front-end processing, we down-sampled the WAV files into 8 kHz, used FaNT Tool to adjust the level of the speech signal [3] and took ITU-tool to filter the WAV file [4]. The commands were

```
filter_add_noise -d -f g712 -i ${in_file}
   -o ${out_file} -s 10 -e file.log -l -20
```

and

```
filter PSO ${in_file} ${out_file}
```

2.3 Feature extraction

We extracted 7-MFCC coefficients (including C0) concatenated with SDC 7-1-3-7, which resulted in 56 coefficients coefficients per frame.

2.4 Statistical modeling

We trained a UBM with 1024 mixtures. The total variability models were built in the same way as depicted in the works [1]. The length normalization was applied. The ivec-LDA system took the normalized ivectors train the LDA matrix. The rank of LDA matrix is 20. The ivec-SVM system took the normalized ivectors train the SVM models.

2.5 Score fusion and calibration

We used focal toolkit to perform score fusion and calibration. The commands were

```
[alpha,beta] =
    train_nary_llr_fusion({ilda_train,isvm_train},classf_train);
discr_fusion_test =
    apply_nary_lin_fusion({ilda_eval,ilda_eval},alpha,beta);
```

3 Contrast 1 system

The contrast 1 system is the raw score of ivec-LDA system. We would like to know the performance of this system without any backend processing.

4 Contrast 2 system

The contrast 1 system is the raw score of ivec-SVM system. We would like to know the performance of this system without any backend processing.

5 Processing time

The CPUs are Intel(R) Xeon(R) Processor E5620 2.40GHz, 12MB cache, 5.86 GT/s QPI, Turbo, HT $\times$ 2. The memory is 16 GB.

The whole system was finished in 48 hours. During the Front-end processing phase, we adopted the public tools and did not calculate it. After the Front-end processing, the RT factor during the recognition phase is ≤ 0.01.

References

1. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, may 2011.
2. Speech database created for the Albayzion 2012 Language Recognition Evaluation. Produced by the Software Technologies Working Group (GTTS), University of the Basque Country UPV/EHT, Spain.
3. G. Hirsch, "Fant", http://dnt.kr.hsnr.de/download.html
4. ITU-T, ITU-T Software Tool Library 2009 Users Manual, ITU-T Users Group on Software Tools, ITU, Geneva, Switzerland, November 2009/

The GTH-LID System for the Albayzin LRE12 Evaluation

Luis Fernando D'Haro, Ricardo Córdoba

Speech Technology Group – Dpto. de Ing. Electrónica – E.T.S.I. de Telecomunicación - Universidad Politécnica de Madrid. Ciudad Universitaria s/n 28040, Madrid, Spain
{lfdharo,cordoba}@die.upm.es

Abstract.
This paper contains a description of the data-sets, systems and fusion alternatives developed by the Speech Technology Group (GTH) for the Albayzin 2012 Language Recognition Evaluation for the 4 conditions: plenty-closed condition (core), plenty-open, empty-closed, and empty-open. In all cases, the primary system is the fusion of three different i-vector based systems: one acoustic system, a phonotactic system using trigrams of phone-posteriorgram counts, and another acoustic system based on RPLP features instead of the traditional MFCC features. For each plenty condition, a contrastive system was also included where the RPLP features or MFCC features where replaced by a different system based on using glottal source features. We provide results for the plenty conditions using the proposed metrics for the evaluation (i.e. Fact, Fdis, and Fcal), as well the known Cavg metric used on NIST evaluations.

Keywords: Language Recognition, Phonotactic system, iVectors, RPLP, GlottHMM

1 Introduction

The goal of this paper is to describe the GTH system for the Albayzin 2012 LRE task. Our primary submission includes three systems:

1. Acoustic system based on MFCC + SDC features and RASTA, iVectors
2. Phonotactic system based on trigram Posteriorgram Counts, iVectors
3. Acoustic system based on RPLP + SDC features and RASTA, iVectors

Fig. 1 shows a block diagram of the submitted system. Detailed information about each system, as well as the calibration and fusion process will be provided throughout the paper. We also submitted a contrastive system for the plenty conditions based on using glottal source features (called GlottHMM-iVector, see section 6). In this case, we replaced system 3 by a system based on using glottal source information and iVectors.

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

As many of current state-of-the-art systems, our three systems make extensive use of sub-space projections in the form of iVectors [1] combining different, but complementary, kind of information.

The paper is organized as follows: Section 2 describes the data-sets used for training, development and test prior to the final evaluation. Section 2.2 explains the acoustic system, section 4 the phonotactic system, section 5 describes the RPLP system, and section 6 the Glottal source based system. Finally, section 7 covers the fusion and calibration results.

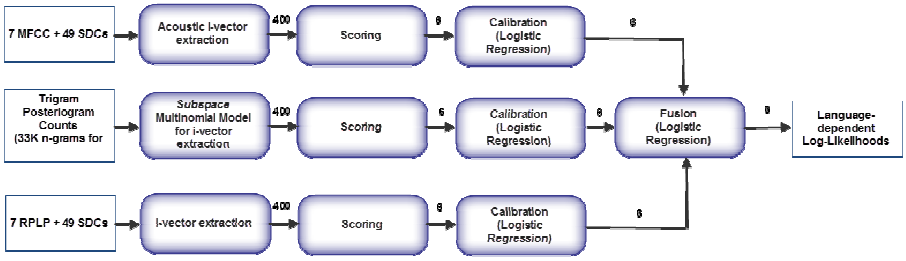


Fig. 1. Block diagram of the primary submitted system.

2 Data description

2.1 Plenty conditions

Table 1 shows the statistics of the number of files used in our setup for training, development and test in both plenty conditions. For the three systems reported in this paper we have always used the same file sets.

Plenty	Closed			Open		
	Train	Dev.	Test	Train	Dev.	Test
No. Files	4656	458	457	5265	725	725
No. Langs	6			7		
No. of clean files	3060	N.A	N.A	3060	N.A	N.A
No. of Noisy files	1596	N.A	N.A	1596	N.A	N.A

Table 1. Statistics for the training, development and test set for the plenty conditions

We have divided the original development set into two subsets with a similar language distribution. We did not apply any k-fold strategy. The first one is the "Dev." set used to calibrate the system, and the second one is the "Test" set, which we have used to obtain the results presented in the paper.

For the final evaluation, we have added the "Test" set to the Train set to have more training data, calibrating the system with the "Dev." set.

2.2 Empty conditions

Table 2 shows the statistics of the number of files used in our setup for training, development and test in both empty conditions.

Empty	Closed			Open		
	Train	Dev.	Test	Train	Dev.	Test
No. Files	-	304	305	-	571	571
Our experiments	7400 (1)	304	305	10141 (2)	571	571
No. Langs	4			5		

Table 2. Statistics for the training, development and test set for the empty conditions

As we did not have any training data for the 4 new languages, we have reused the training set from the plenty conditions and merged them with the development data available duplicated three times to give it more relevance. We did not apply any adaptation technique to the models from the plenty conditions.

In detail, for (1) we have merged the data from the plenty closed (PC) training data with the PC Development data and 3 times the empty closed (EC) Development data, giving a total of 7400 training examples. In the same way, for (2), we have merged the data from plenty open (PO) training data with the PO Development data and 3 times the empty open (EO) Development data, giving a total of 10141 training examples.

As for the plenty conditions, for the final evaluation we have also added the "Test" set to the Train set to have more training data, calibrating the system with the "Dev." set. For training the logistic regression classifier for the EC condition, we have unified the "Dev." and "Test" sets but the calibration was done only on the "Dev." Set. For the EO condition we have used 10141 files for training the LR classifier and the calibration was done on the 571 files.

3 Acoustic system based on MFCC + SDC features and RASTA, iVectors

In this section, we provide a brief summary of the acoustic feature extraction and UBM training used for training our acoustic system.

The first step in our system is to extract the feature vectors from variable duration segments of recorded speech. In order to do this, we first parameterize each file using SPRO5[1] extracting 12 MFCC coefficients (including C0) from 24 Mel filter banks plus the energy for each frame. Finally, Cepstral mean and variance normalization is applied to the feature vectors for each file.

[1] https://gforge.inria.fr/projects/spro

The Voice Activity Detector (VAD) used for all the systems is the output from the the BUT Hungarian phone recognizer[2]. Then, we suppressed all segments marked as silence or noise in the output. After discarding the silence segments, every 10 ms speech frame was mapped to a 56-dimensional feature vector. The feature vector is the concatenation of SDC features [3] using the common 7-1-3-7 configuration and stacking them with the first 7 MFCC coefficients out from the 12 MFCCs. Finally, a RASTA filter was applied in order to reduce short-term noise variations in each frequency subband. No Vocal-Tract Length Normalization (VTLN) was applied.

Then, we train a language-independent GMM, a.k.a universal background model (UBM), through five iterations of the EM-algorithm and using all the acoustic feature vectors coming from all the 6 (or 7 for the open condition) languages that appeared in the training set.

3.1 Acoustic iVectors

Currently one of the main techniques used for speaker recognition [2] and language recognition [4][5] is the iVectors technique. In this framework, the language and channel-dependent GMM supervector M can be modeled as:

$$M = m + Tw \quad (1)$$

Where m is the UBM GMM mean supervector, T is the total variability matrix (i.e. the iVector extractor) and w is a standard normal distributed vector of size M (i.e. iVector). The main advantage of w is that it maps most of the relevant information from the variable-length audio file to a fixed-length and small dimensional vector.

Finally, the iVectors are normalized by first subtracting the mean of all the training iVectors and then dividing them by its corresponding norm.

3.2 Results

Table 3 shows the results for the systems presented.

Condition	System	Closed		Open	
		Fact	Cavgx100	Fact	Cavgx100
Plenty	300iv, 512 Gauss.	0.176646	9.08	0.202484	10.54
	400iv, 512 Gauss.	0.172519	9.25	-	-
	400iv, 1024Gauss.	0.172484	9.19	0.196182	10.37
Empty	400iv, 64 Gauss.	0.075818	0.43	0.092105	3.32
	400iv, 128 Gauss.	0.092308	0.83	0.115086	4.24
	400iv, 256 Gauss.	0.109475	1.57	0.149377	5.70
	400iv, 512 Gauss.	0.120738	2.10	-	-

Table 3. Reported results for the acoustic MFCC system on the test set

[2] http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

As we can see, in the plenty conditions, 400 iVectors provide better results than 300, and there is little difference between 512 and 1024 Gaussians, so we will use the 512 Gaussians, as it is faster and probably more robust for unseen test examples.

For the empty conditions, as there is little data, best results are obtained for 64 Gaussians, so this is the system used in the evaluation.

4 Phonotactic system based on trigram Posteriorgram Counts, iVectors

4.1 System description

In this case we have used a novel approach to phonotactic LID reported in [10], where instead of using soft-counts based on phoneme lattices, we use posteriorgrams to obtain n-gram counts. In this approach, the high-dimensional vectors of counts are reduced to low-dimensional units for which we adapted the commonly used technique iVectors. The reduction is based on multinomial subspace modeling and is designed to work in the total-variability space. In comparison with the other techniques based on soft-counts, the new technique provides better results, reduces the problems due to sparse counts, and avoids the process of using pruning techniques when creating the lattices. Previous reported results for the NIST 2009 LRE data-set showed better results compared to a system based on using soft-counts, and with very good results when fused with an acoustic i-vector LID system. For this reason, we decided to use this technique in this evaluation and check its behavior on a different database. Next, we briefly describe the proposed technique.

Feature extraction.

Fig. 2 shows the process of creating the vector of posteriorgram-based n-gram counts. In the figure, we consider the bigram counts for simplicity, but in our system we used trigrams. The process can be divided into four main steps:

1. The first step is to tokenize speech by the means of running a phone-recognizer that, for each frame, provides the posterior probabilities of the phone occurrences. In our experiments, we used the BUT Hungarian phone recognizer.
2. The second step is to sum up and average the posterior probabilities for the frames that are considered to be within the same phoneme unit (A, B, C in the Figure). The phone boundaries are obtained by running Viterbi decoding on the posteriorgram. The averaged posterior provides a good de-correlation and smoothness for the resulting matrix that we call *averaged posteriorgram*.
3. The third step is to create the *joint-posteriorgram* – a sequence of matrices of joint probabilities for the *n* consecutive frames. Here, we take the averaged posteriorgram of each frame and we do the outer product with the posteriorgram of the previous frame. Then, the process is repeated for all the phone-grams considering the *n*-1 phone-gram history.

4. The final step is to sum up all frames (matrices) of the joint-posteriorgram. This way, we create a matrix of *n*-gram counts that is converted into a 1xD vector (where D is the total number of possible *n*-grams, in the case of trigrams is 35937, 33^3, 33 phonemes and order 3) and then used as a feature file for training the iVectors using Subspace Multinomial Models.

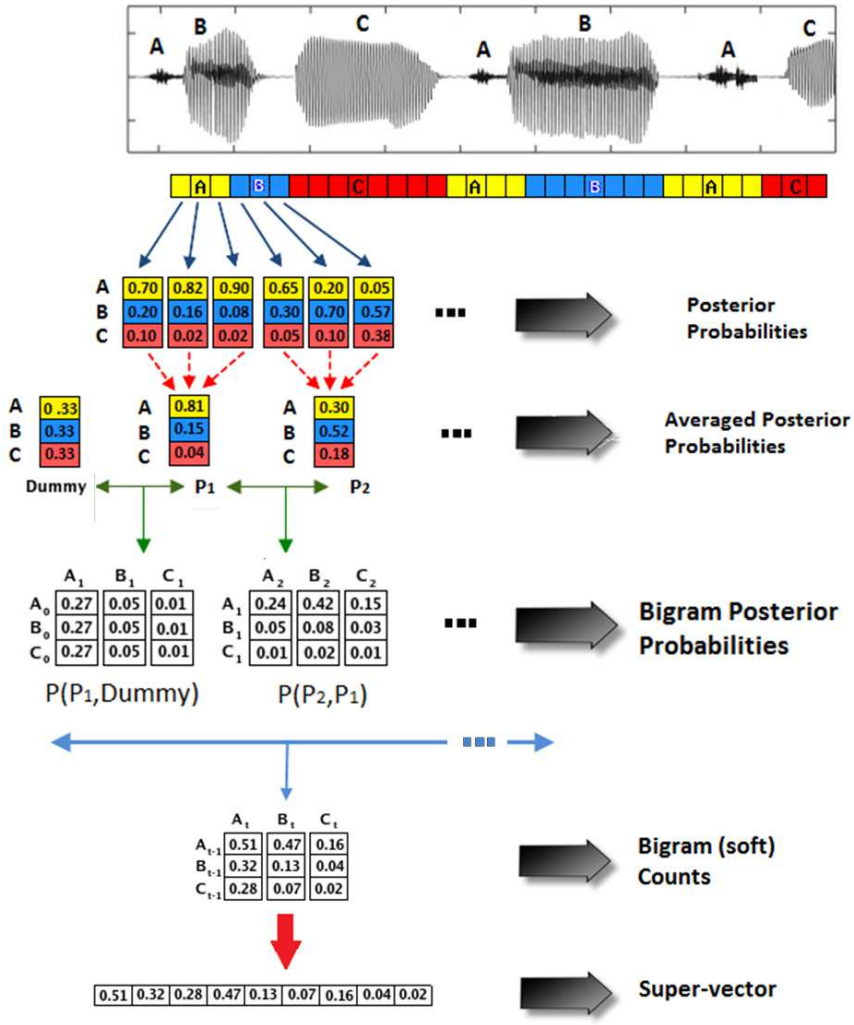


Fig. 2. Procedure to generate posteriorgram counts features

Subspace Multinomial Models.

The goal of the Subspace Multinomial Model is to model the discrete representation of the posteriorgram counts created in the previous step. Thanks to the Subspace Multinomial Models we can train low dimensional vectors of coordinates in the total

variability subspace, i.e. iVectors, and then use these iVectors as a feature vector input for training a discriminative LID classifier. Next, we briefly describe the process for training the subspace multinomial models. For further details please refer to [11] and [12].

The log-likelihood of data D for a multinomial model with C discrete classes is determined by model parameters φ and sufficient statistics γ, representing the occupation counts of classes for all N utterances in D:

$$\log p(D) = \sum_{n=1}^{N} \sum_{c=1}^{C} \gamma_{nc} \log \varphi_{nc} \quad (2)$$

Where γ_{nc} is the occupation count for class c and utterance n and φ_{nc} are probabilities of (utterance dependent) multinomial distribution, defined by a subspace model according:

$$\varphi_{nc} = \frac{\exp(m_c + t_c w_n)}{\sum_i^C \exp(m_i + t_i w_n)} \quad (3)$$

Where t_c is the c-th row of subspace matrix T and w_n is an r dimensional column vector (i-vector) representing language and channel of utterance n.

For training the iVectors, we have followed the algorithm reported in [11] with slight modifications in order to iterate several times the estimation and maximization steps (further details can be found in [10]). Finally, in our experiments, we have considered a set of 1089 multinomial models when using trigrams (i.e. considering all the possible number of bigram histories, 33x33, using 33 phones for the Hungarian recognizer).

Finally, it is important to mention that for the empty condition we have implemented the following algorithm. First, we obtain the T matrix by using the created training set described in section 2.2 together with the development set and applying two epochs and two iterations for the EM iVectors extraction process. Then, the new T matrix is used to extract the final iVectors for all sets.

4.2 Results

Condition	System	Closed		Open	
		Fact	Cavg x100	Fact	Cavg x100
Plenty	400iv_Trigrams	0.138718	9.43	0.163411	10.37
Empty	400iv_Trigrams	0.037714	0.17	0.047180	2.40

Table 4. Reported results for the phonotactic system on the test set

5 Acoustic system based on RPLP + SDC features and RASTA, iVectors

5.1 System description

The goal of this system was to introduce a new set of features which could be more robust to noise. In this case, we decided to use the RPLP (Revised PLP) features used in [13] and proposed in [14]. These features can be seen as a hybrid approach between MFCC and PLP, combining the advantages of both. **Fig. 3** shows the modules needed to calculate the traditional features (MFCC and PLP) compared to these new RPLP. As we can see, the procedure to calculate them is very similar to the MFCC computation but the DCT-based transformation is replaced by an auto-regressive (AR) modeling with additional decreasing of spectral dynamics using the INtensity-TO-LouDness (IN2LD) factor (i.e. the power-law) introduced during the PLP calculation. In [13] good improvements were found for ASR recognition in comparison with the standard features. One important contribution of this method is that it performs a double suppression of spectral dynamics before calculating the cepstral coefficients (LPC), and with less effect on the accuracy when modifying the number of FB bands, shape, and non-linearity scaling.

Finally, we apply a RASTA filter to these coefficients and then we calculate the SDC parameters for the first 7 RPLP, obtaining a final vector of dimension 56, using the common 7-1-3-7 configuration.

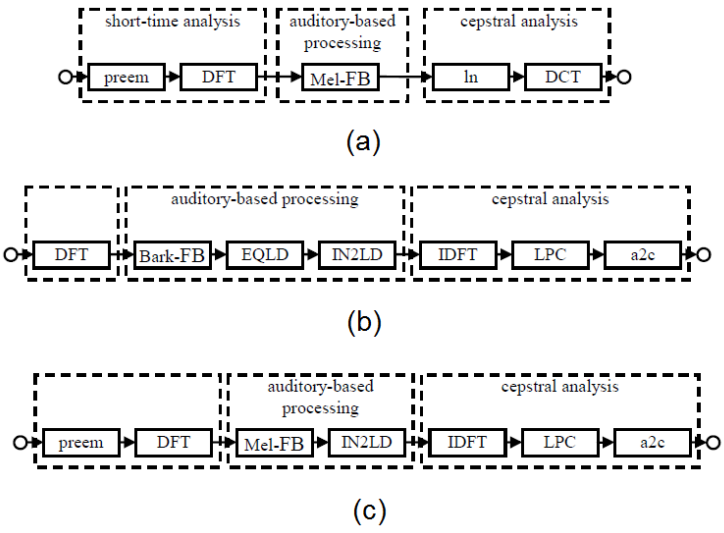


Fig. 3. Module sequence for calculating (a) MFCC features, (b) PLP features, and (c) RPLP features

5.2 Results

Table 5 shows the results for the systems presented.

Condition		Closed		Open	
		Fact	**Cavgx100**	**Fact**	**Cavgx100**
Plenty	400iv, 512 Gauss.	0.159279	7.62	0.173536	10.20
	400iv, 1024Gauss.	0.156287	7.60	0.172224	10.27
Empty	400iv, 64 Gauss.	0.040305	0.099	0.054902	1.84
	400iv, 128 Gauss.	0.041599	0.099	0.053973	1.39
	400iv, 256 Gauss.	0.038545	0.047	0.050978	1.27

Table 5. Reported results for the acoustic RPLP system on the test set

As we can see, in the plenty conditions, there is little difference between 512 and 1024 Gaussians, so we will use the 512 Gaussians, as it is faster and probably more robust for unseen test examples. It is important to see that results are better than those obtained with MFCC (see **Table 3**). For example, for the Plenty-closed condition the relative improvement is 6.42%.

For the empty conditions, we have found much better results than using MFCC (see Section 3), and we could even use 256 Gaussians.

6 GlottHMM-iVector: Glottal source based system

6.1 System description

The goal of this system was to check the viability of using glottal source features for language recognition based on the good results reported by [6] on speaking style identification in the Ircam database, and by [7] for classifying expressive speech. In the former, the use of only prosodic information (i.e pitch and rhythm) provided identification rates of about 74%; for the later, a 95% for styled speech and 82% for emotional speech on a different database.

GlottHMM [8] is a vocoding technique that was recently developed for parametric speech synthesis. It is based on decomposing speech into the glottal source and vocal tract through glottal inverse filtering. In our system we have used GlottHMM to extract the F0 and the Harmonics to Noise Ratio (HNR) of the glottal source. For this system, the F0 information is used as VAD to separate between voiced and unvoiced frames. HNR is evaluated based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and inter-harmonic valleys) and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale. Choosing the same selected parameters reported in [7], our feature vector was made by concatenating the: F0 and the five HNR coefficients, and then calculating the SDCs coefficients on the same way as for the acoustic systems in our primary submission. Then, we used the iVectors technique following the same approach as for the Acoustic system.

In addition, and in an effort to try to select the same parameters reported in [7], we also included into our feature vector the selected parameters of the vocal tract obtained using Line Spectral Frequency (LSF) with a vector of length 30, the selected parameters of the spectral tilt of the glottal source modeled using LSFs (with 10 LSFs), and the Normalized Amplitude Quotient (NAQ) [9] and the magnitude differences between the 10 first harmonics of the voice source.

Finally, it is important to mention that in the process of extracting the features using GlottHMM, it uses a high-pass filter whose purpose is to reduce low-frequency components that may cause large fluctuations in the estimated glottal flow signal. The tool includes two files with the coefficients required to process files sampled at 16KHz or 44KHz, that can be specified through the configuration file. In addition, files with other sampling rates can be processed thanks to a simple Matlab script for designing FIR filters that is provided with the package. In our case, since the original files provided by the organizers from the Kalaka3 database are sampled at 16 KHz we used the default corresponding coefficients file provided by the tool.

6.2 Results

Table 6 shows the results for the systems presented.

Condition	System	Closed		Open	
		Fact	**Cavgx100**	**Fact**	**Cavgx100**
Plenty	400iv, 64 Gauss.	0.720873	33.11	0.740958	34.02
	400iv, 128 Gauss.	0.633388	28.61	0.715139	32.26
	400iv, 256 Gauss.	0.673944	30.67	0.710156	32.94
	400iv, 512 Gauss.	0.668717	30.81	0.718101	32.64
Empty	400iv, 8 Gauss.	0.075131	2.38	0.163136	7.99
	400iv, 16 Gauss.	0.082595	2.29	0.162338	7.96
	400iv, 32 Gauss.	0.113325	4.19	0.210829	11.33
	400iv, 64 Gauss	0.160465	5.06	0.304118	15.28
	400iv, 128 Gauss.	0.211179	8.18	0.472886	23.32

Table 6. Reported results for the GlottHMM system on the test set

As this system provided worse results it was only used in the contrastive systems for all conditions. In the plenty conditions, best results are obtained using 128 Gaussians. In the empty conditions, results clearly improved using very few Gaussians. We finally decided to use only 16 Gaussians for the contrastive system using this features.

7 Classifier and Calibration Back-end

As classifier for our three iVectors systems, we have used a Multiclass logistic regression which generates a different classifier for each language to recognize. Then, these classifiers are used to generate scores for the files in our development and test sets. For calibration and fusion, a Gaussian Back-end followed by a Discriminative

Multi-Class Logistic Regression is used to post-process the scores obtained before. Previously, the input vectors were conditioned by within-class covariance normalization (WCCN).

Regarding our calibration and fusion module, it is mainly based on the FoCal toolkit[3], so we did not use the tools provided by the Albayzin organizers.

Table 7 shows the results for the systems presented.

System 1	System 2	System 3	Closed		Open	
			Fact	Cavgx100	Fact	Cavgx100
MFCC-512G	Phon-1089G	RPLP-512G	0.069258	4.16	0.080184	5.77
MFCC-512G	Phon-1089G	Glot-128G	0.071393	4.16	-	-
RPLP-512G	Phon-1089G	Glot-512G	-	-	0.079517	5.37

Table 7. Reported results for the fusion of all systems for the plenty condition

The first line corresponds to the primary system submitted to the evaluation and the second and third lines correspond to the contrastive systems using GlotHMM. We can see that the contrastive system in the Open condition is slightly better than the primary one, even though the results for GlotHMM and clearly worse than for the MFCC system. Also, the GlotHMM with 512 Gaussians provided slightly better results than with 128 Gaussians. However, we decided to keep the primary system using MFCC.

8 Summary

As a summary, here is a list of the systems presented for the evaluation:

Condition	System 1	System 2	System 3
Plenty- closed-primary	MFCC-512G	Phon-1089G	RPLP-512G
Plenty-closed-contrastive1	MFCC-512G	Phon-1089G	Glot-128G
Plenty-open-primary	MFCC-512G	Phon-1089G	RPLP-512G
Plenty-open-contrastive1	RPLP-512G	Phon-1089G	Glot-512G
Empty-primary	MFCC-64G	Phon-1089G	RPLP-256G
Empty-contrastive1	RPLP-256G	Phon-1089G	Glot-16G
Empty-open-primary	MFCC-64G	Phon-1089G	RPLP-256G
Empty-open-contrastive1	RPLP-256G	Phon-1089G	Glot-16G

Table 8. Summary of all systems presented to the evaluation

[3] https://sites.google.com/site/nikobrummer/focal

9 Acknowledgment

This work has been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects, and the European project Simple4All (under grant agreement No. 287678).

10 Bibliography

1. Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel, “Support vector machines versus fast scoring in the low dimensional total variability space for speaker verification,” in Proc. of Interspeech 2009, Brighton, UK.
2. Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis For Speaker Verification", IEEE Transactions on Audio, Speech and Language Processing, Vol. 19 (4), May 2011.
3. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2002, pp. 89–92.
4. D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space", in Proc. of Interspeech 2011.
5. N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via Ivectors and Dimensionality Reduction", in Proc. of Interspeech 2011.
6. Nicolas Obin, “MeLos: Analysis and Modelling of Speech Prosody and Speaking Style”, PhD, Ircam-UPMC, Paris, 2011.
7. Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Tuomo Raitio, Nicolas Obin, Paavo Alku, Junichi Yamagishi, Juan M Montero. "Towards Glottal Source Controllability in Expressive Speech Synthesis", in Proc. of Interspeech 2012.
8. Raitio, T. and Suni, A. and Yamagishi, J. and Pulakka, H. and Nurminen, J. and Vainio, M. and Alku, P., ”HMM-based speech synthesis utilizing glottal inverse filtering”, Audio, Speech, and Language Processing, IEEE Transactions on, 9:153-165, IEEE, 2011.
9. P. Alku, T. Bäckström, and E. Vilkman, “Normalized amplitude quotient for parameterization of the glottal flow”, J. Acoust. Soc. Amer., vol. 112, no. 2, pp. 701-710, 2002.
10. Luis Fernando D’Haro, Ondrej Glembek, Oldrich Plchot, Pavel Matejka, Mehdi Soufifar, Ricardo Cordoba, Jan Cernocký. “Phonotactic Language Recognition using i-vectors and Phoneme Posteriorgram Counts”, in Proc. of Interspeech 2012.
11. Kockmann, et al, 2010. “Prosodic speaker verification using subspace multinomial models with intersession compensation,” in Proc. of ICSPL, Makuhari, Chiba, Japan, 2010.
12. D. Povey, Lukas Burget et. al, 2011. "The Subspace Gaussian Mixture Model– a Structured Model for Speech Recognition", Computer Speech and Language, 25(2), pp. 404-439
13. Rajnoha, J., and Pollák, P. 2011. “ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness”. Radionegineering, Vol. 20, No. 1, April 2011, pp. 74-84.
14. Hönig, F., Stemmer, G., Hacker, C., Brugnara, F. "Revising Perceptual Linear Prediction (PLP)". In Eurospeech 2005, p. 2997-3000.

The L^{2}F Language Recognition System for Albayzin 2012 Evaluation

Alberto Abad

L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa,
alberto@l2f.inesc-id.pt,
WWW home page: http://www.l2f.inesc-id.pt

Abstract. This document presents a description of INESC-ID's Spoken Language Systems Laboratory (L^{2}F) systems submitted to the Albayzin 2012 Language Recognition evaluation. The submitted systems differ on the number of sub-systems selected for fusion and the back-end configuration. The basic set of sub-systems considered are four conventional phonotactic sub-systems based on n-gram modelling of phoneme sequences, four additional phonotactic sub-systems based on SVM discriminative modelling of expected phone counts extracted from lattices, and an i-vector based sub-system with linear generative classifiers. Similarly to the L^{2}F Albayzin 2010 system, individual language models for clean and noisy conditions have been trained for each target language of the *Plenty of Training* condition. The L^{2}F *primary* system exploits Gaussian back-ends for each sub-system and linear logistic regression fusion of k sub-systems, selected automatically following a non-exhaustive fast greedy search method to find the best (sub-optimal) combination. This search process and the determination of the back-end parameters is performed per evaluation condition. Additionally, three contrastive systems have been developed. Language detection results have been submitted for all the evaluation conditions for every system.

Keywords: language recognition, Albayzin evaluations

1 Introduction

The "Red Temática en Tecnologías del Habla" (RTTH) has organised in the recent years a series of evaluations - so called Albayzin evaluations - in some relevant speech processing topics devoted to encourage language research activities on the four official languages of Spain. Similar to the well-known NIST Language Recognition Evaluation, a series of Language Recognition (LR) tasks have been proposed in 2008 and 2010. In the new Albayzin 2012 Language Recognition Evaluation there are significant novelties with respect to the previous editions. In contrast to previous campaigns, this year test data is considerably more challenging and consists of audios extracted from Youtube videos. Moreover, two different evaluation conditions have been proposed: *Plenty of Training* and *Empty Training*. For the *Plenty* condition, training data is provided for the target languages (Castilian, Catalan, Basque, Galician, Portuguese and English) and it

can be used to train language models like in previous evaluation editions. In the new *Empty* condition, training data for the target languages (French, German, Greek and Italian) is not provided. In both cases, it is not allowed to use additional data from external sources for the development of the LR systems. Moreover, like in previous campaigns, *closed-set* and *open-set* conditions are also defined, resulting in a total of four possible evaluation conditions: *Plenty-Closed* (PC), *Plenty-Open* (PO), *Empty-Closed* (EC) and *Empty-Open* (EO). Detailed information on the evaluation campaign can be found in the evaluation plan document [1].

This document presents the LR systems developed by INESC-ID's Spoken Language Systems Laboratory (L^2F) for the Albayzin 2012 campaign. LR approaches can generally be classified according to the kind of source of information that they rely on. The most successful systems are based on the exploitation of the acoustic phonetics, that is the acoustic characteristics of each language, or the phonotactics which are the rules that govern the phone combinations in a language. Usually, the combination of different sources of knowledge and systems of different characteristics tends to provide increased language recognition performances [2]. For this evaluation, nine sub-systems have been developed: four phonotactic based on Phone Recognition and Language Modelling (PRLM) [3], four phonotactic based on Phone Recognisers followed by support vector machine modelling (PRSVM) [4] and an i-vector [10] based language recognition system similar to the one in [5] that makes use of single mixture Gaussian distributions for language modelling. A primary and three contrastive systems have been submitted, which differ in the number of employed sub-systems and in the back-end strategy followed. All the submitted LR systems implement Gaussian back-ends followed by linear logistic regression fusion. The *primary* system follows a greedy search strategy to find the best combination of sub-systems per evaluation condition like in [6]. The *contrastive1* system follows the same subsystem selection approach, but applies zt-norm to the phonotactic scores. The *contrastive2* system consists of the fusion of the four PRSVM sub-systems and the i-vector sub-system. The *contrastive3* system fuses the nine sub-systems.

In next section 2 a brief description of some commonalities of the subsystems developed (see Section 2.1) is provided, together with details of each one of the nine individual sub-systems: the PRLM-LR, the PRSVM-LR and the iVECTOR-LR sub-systems are described in sections 2.2, 2.3 and 2.4, respectively. Finally, details about the back-end and fusion and about the four submitted systems are provided in section 3.

2 LR sub-system description

2.1 Sub-system commonalities

Data Pre-processing Training data provided for the evaluation consist of two sets of clean speech (around 86 hours) and noisy speech (around 22 hours) broadcast data for each one of the 6 target languages considered in the *Plenty*-training condition: Basque, Catalan, English, Galician, Portuguese and Spanish.

The training data was pre-processed to segment long data files into a set of homogeneous reduced length speech segments. In order to generate these homogenous segments, we applied our segmentation module [7], including speech-non-speech (SNS) segmentation, background classification, channel classification, gender classification and speaker clustering. After this segmentation process, we selected for each target language 5 hours of clean speech (segments with minimum duration of 15 seconds and maximum duration of 40 seconds), and 1.5 hours of noisy speech (segments with minimum duration of 10 seconds and maximum duration of 40 seconds). Table 1 shows the amount of selected segments and the average duration per segment in seconds for each target language and type of speech. After the segmentation process, all training segments are down-sampled to 8kHz sampling rate.

On the other hand, the development and evaluation data sets consist of audio extracted from Youtube videos. In this case, during the development of the systems we experimented two alternative pre-processing strategies. First, we considered removing non-speech segments detected with our segmentation module to produce a "cleaned" version of the development data set. Second, we segmented each development file in shorter speech homogeneous segments that were independently processed to obtain several language recognition scores per file. Then, we experimented some simple strategies to generate a single score. None of these two mentioned strategies provided any observable improvement with respect to the use of the whole unprocessed test segment. Consequently, it was decided to not apply any additional pre-processing to the development and evaluation data sets, besides downsampling to 8kHz.

	clean		noisy	
	#segm	mean dur. [sec]	#segm	mean dur. [sec]
Basque	778	23.1	291	17.6
Catalan	772	23.3	299	18.1
English	736	24.5	284	19.0
Galician	783	23.0	310	17.5
Portuguese	770	23.4	299	18.1
Spanish	753	23.9	282	19.2

Table 1. Training data segmentation for each target language and speech type.

Target Language Modelling One of the particularities shared among all the developed sub-systems is that a separate target language (of the *Plenty*-training condition) model was trained for clean and noisy speech. The two target models of each language are used to obtain two language-dependent scores for each speech test segment. Consequently, for every test segment a vector of 12 scores $\mathbf{x}_i$ is produced by every individual sub-system i.

2.2 PRLM-LR sub-systems

The Phone Recognition followed by Language Modelling (PRLM) systems used for Albayzin 2012 exploit the phonotactic information extracted by four individual tokenizers: European Portuguese (*pt*), Brazilian Portuguese (*bp*), European Spanish (*es*) and American English (*en*). The key aspect of this type of system is the need for robust phonetic classifiers that generally need to be trained with word-level or phonetic level transcriptions. In this case, the tokenizers are MultiLayer Perceptrons (MLP) trained to estimate the posterior probabilities of the different phonemes for a given input speech frame (and its context). For each target language and for each tokenizer a different phonotactic *n-gram* language model is trained. During test, the phonetic sequence of a given speech signal is extracted with the phonetic classifiers and the likelihood of each target language model is evaluated.

Phonetic Tokenizers The tokenization of the speech data is done with the neural networks that are part of our hybrid Automatic Speech Recognition (ASR) system named AUDIMUS [8]. The recognisers combine four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and Advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives). A phone-loop grammar with phoneme minimum duration of three frames is used for phonetic decoding.

The language-dependent MLP networks were trained with different amounts of annotated data. For the *pt* acoustic models, 57 hours of BN down-sampled data and 58 hours of mixed fixed-telephone and mobile-telephone data were used. The *bp* models were trained with around 13 hours of BN down-sampled data. The *es* networks used 36 hours of BN down-sampled data and 21 hours of fixed-telephone data. The *en* system was trained with the HUB-4 96 and HUB-4 97 down-sampled data sets, that contain around 142 hours of TV and Radio Broadcast data.

Each MLP network is characterised by the size of its input layer that depends on the particular parametrization and the frame context size (13 for PLP, PLP-RASTA and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. In this case, only monophone units are modelled, resulting in MLP networks of 41 (39 phonemes +1 silence + 1 respiration) soft-max outputs in the case of *en*, 39 for *pt* (38 phonemes + 1 silence), 40 for *bp* (39 phonemes + 1 silence) and 30 for *es* (29 phonemes + 1 silence).

Phonotactics Modelling For every phonetic tokenizer, the phonotactics of each target language for every type of speech condition (clean and noisy) is

modemed with a 3-gram back-off model, that is smoothened using Witten-Bell discounting. For that purpose the SRILM toolkit has been used[1].

2.3 PRSVM-LR sub-systems

Phone Recognition followed by Support Vector Machine Modelling (PRSVM) systems used for Albayzin 2012 exploit the phonotactic information extracted by the same four tokenizers described above: *pt*, *bp*, *es* and *en*. In contrast to PRLM-LR sub-systems, a recognition lattice is generated for every processed segment, from which the posterior expected n-gram counts are computed. Then, for each target language and for each tokenizer a different phonotactic SVM language model is trained with the counts vectors. During test, vectors of n-gram counts of a given speech signal are computed from the lattices obtained with the automatic phoneme recognisers and used to evaluate each language SVM model.

Phoneme Recognisers Vectors of expected n-gram counts are obtained for each speech segment based on the recognition results of our ASR system described above. Like in PRLM sub-systems, a phone-loop grammar with phoneme minimum duration of three frames is used for lattice generation.

N-gram vector extraction and dimensionality reduction The 'lattice-tool' program from the SRILM toolkit is used to compute the expected n-gram counts (up to 3-grams) of each recognition lattice. This resulting n-gram counts vector is converted to a vector of probabilities (sum 1) and it is normalised by the square root of the average probability vector computed over the whole training data set. The high-dimensionality of the n-gram vectors motivated the use of some sort of dimensionality reduction method. In practice, we applied simple frequency selection [9] with new dimensionality of 10000 elements in the four PRSVM sub-systems (this size was experimentally verified to provide good performance).

Phonotactics Modelling For every phoneme recogniser, phonotactic relations of each training data sub-set are modelled with an L2-regularised support vector classifier using the LibLinear implementation of the libSVM tool[2]. For "clean" and "noisy" SVM language model training, only "clean" and "noisy" background (non-positive) data are used respectively.

2.4 iVECTOR-LR sub-system

Total-variability modelling [10] has rapidly emerged as one of the most powerful approaches to the problem of speaker verification. In this approach, closely

[1] `http://www-speech.sri.com/projects/srilm/`
[2] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

related to the Joint Factor Analysis [11], the speaker and the channel variabilities of the high-dimensional GMM supervector are jointly modelled as a single low-rank total-variability space. The low-dimensionality total variability factors extracted from a given speech segment form a vector, named i-vector, which represents the speech segment in a very compact and efficient way. Thus, the total-variability modelling is used as a factor analysis based front-end extractor. In practice, since the i-vector comprises both speaker and channel variabilities, in the i-vector framework for speaker verification some sort of channel compensation or channel modelling technique usually follows the i-vector extraction process.

The success of i-vector based speaker recognition has motivated the investigation of its application to other related fields, including language recognition [5, 12]. For Albayzin 2011, we have developed an i-vector based language recognition sub-system very similar to the one in [5], where the distribution of i-vectors for each language is modelled with a single Gaussian.

Feature extraction The extracted features are shifted delta cepstra (SDC) [13] of Perceptual Linear Prediction features with log-RelAtive SpecTrAl speech processing (PLP-RASTA). First, 7 PLP-RASTA static features are obtained and mean and variance normalisation is applied in a per segment basis. Then, SDC features (with a 7-1-3-7 configuration) are computed, resulting in a feature vector of 56 components. Finally, low-energy frames detected with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment are removed.

UBM modelling A GMM-UBM of 1024 mixtures has been trained using all the training segments of Table 1. Type of speech was not distinguished and a single UBM was trained with both clean and noisy segments. In total, the number of segments considered are 6330, which corresponds almost to 22.5 hours of speech (after the low-energy removal process of the feature extraction).

Total variability and i-vector extraction The total variability factor matrix ($\mathbf{T}$) was estimated according to [14]. The dimension of the total variability subspace was fixed to 400. Zero and first-order sufficient statistics of the training sub-sets described in Table 1 were used for training $\mathbf{T}$. 10 EM iterations were applied, in the first 7 iterations only ML estimation updates were applied, while in the last 3 EM iterations both ML and minimum divergence update were applied. The covariance matrix was not updated in any of the EM iterations.

The estimated $\mathbf{T}$ matrix is used for extraction of the total variability factors of the processing speech segments as described in [14]. Finally, the resulting factor vectors are normalised to be of unit length, which we will refer as i-vectors.

Language modelling and scoring Like in [5], all the extracted i-vectors from a data sub-set of Table 1 are used to train a single mixture Gaussian distribution

with full covariance matrix shared across different training sub-sets. For a given test i-vector, each Gaussian model is evaluated and log-likelihood scores are obtained.

3 The L^{2}F submitted systems

3.1 Back-end configuration and calibration

Linear Gaussian Back-End A linear Gaussian Back-End (GBE) follows every single sub-system to transform the 12 elements score-vector $\mathbf{x}_i$ (see section 2.1) to a n elements log-likelihood vector $\mathbf{s}_i$, where n equals the number of target languages in closed evaluation conditionS and equals the number of target languages plus 1 out-of-set log-likelihood in open set conditions:

$$\mathbf{s}_i = \mathbf{A}_i \mathbf{x}_i + \mathbf{o}_i \tag{1}$$

where $\mathbf{A}_i$ is the transformation matrix for system i and $\mathbf{o}_i$ is the offset vector.

Linear logistic regression (LLR) Linear logistic regression (LLR) is used to fuse the log-likelihood outputs generated by the linear GBEs of the selected sub-systems to produce fused log-likelihoods $\mathbf{l}$:

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b} \tag{2}$$

where α_i is the weight for sub-system i and $\mathbf{b}$ is the language-dependent shift.

During the development of the L^{2}F systems, the GBEs and the LLR fusion parameters were trained and evaluated with the development data set using a sort of 2-fold cross-validation [6]: development data is randomly split in two halves, one for parameter estimation and the other for assessment. This process is repeated using 10 different random partitions and the mean and variance of the systems' performance can be computed. For the final submission, no partition of the data was made and all the development data was used to simultaneously calibrate the GBEs and the LLR fusion. Different GBE and LLR fusion parameters have been trained for each one of the four evaluation conditions. Calibration was carried out using the FoCal Multi-class Toolkit[3].

3.2 Primary System (primary)

The L^{2}F *primary* system consists of multi-class fusion of a selected set of sub-systems. For a given test segment, the outcome of the fusion is a likelihood vector $\mathbf{l}$ of n-elements, one for each target language (plus 1 for the out-of-set in the open-set condition). The selection of sub-systems is done following an

[3] `http://niko.brummer.googlepages.com/focalmulticlass`

incremental search process using the development data. First, it is found the best single sub-system $[i]$, then the best combination of 2 sub-systems $[i, j]$ with sub-system i, then the best combination of three sub-systems with the best pair previously found, and so on. Finally, the combination of k sub-systems with the lowest minimum performance cost is selected. The search process, and consequently the selection of sub-systems, was done for each evaluation condition independently. Table 2 shows the selected sub-systems of the *primary* system for each evaluation condition. The minimum number of selected sub-systems is 4 for the PC condition and the maximum is 6 for the *Empty* training conditions. In this case, the sub-systems selected for the PC condition are always present in the other conditions. Notice, however, that there is not any restriction to make this happen and, moreover, that the order of selection may not be the same for all the conditions. An interesting observation is that PRLM and PRSVM sub-systems based on the same phonetic classifier are sometimes selected before than other phonotactic systems exploiting different phonetic recognisers. This observation may suggest that there may be some residual complementary information in n-gram and expected counts based phonotactic approaches.

3.3 First Contrastive System (contrastive1)

The L^{2}F *contrastive1* system follows the same sub-system search approach than the *primary* system with a slightly different back-end configuration. Concretely, zt-norm score normalisation is applied to each sub-system before the application of the GBE. In practice, we observed a generalised improvement of the individual sub-systems using score normalisation, with the exception of the iVECTOR-LR sub-system. Consequently, the *contrastive1* system back-end configuration applies zt-norm only to the phonotactic-based sub-systems. Table 2 details the set of sub-systems that form the *contrastive1* system per evaluation condition. In this case, the minimum number of selected sub-systems is 5. Moreover, all the sub-systems selected in the PC condition are not present in all the other conditions, in contrast to the *primary* system. Again, some phonotactic n-gram and expected counts based sub-systems using the same phonetic decoder are selected.

3.4 Second Contrastive System (contrastive2)

The L^{2}F *contrastive2* system consists of the fusion of a fixed set-up of sub-systems for all the evaluation conditions, which are the four PPRSVM-LR sub-systems plus the iVECTOR-LR one. This submission is very similar to the L^{2}F system submitted to NIST LRE2011 [15] (the NIST system incorporates an additional Gaussian supervector based sub-system [16]). Score normalisation is not applied to any of the sub-systems that form the *contrastive2* submission.

3.5 Third Contrastive System (contrastive3)

The L^{2}F *contrastive3* system is the result of the fusion of the nine developed sub-systems. No score normalisation is applied.

System	Cond.	Selected sub-systems
Primary	PC	PRLM-es, PRSVM-bp, PRSVM-es, iVECTOR
	PO	PRLM-es, PRSVM-bp, PRSVM-en, PRSVM-es, iVECTOR
	EC	PRLM-en, PRLM-es, PRSVM-bp, PRSVM-en, PRSVM-es, iVECTOR
	EO	PRLM-es, PRSVM-bp, PRSVM-en, PRSVM-es, PRSVM-pt, iVECTOR
Contrastive1	PC	PRLM-bp, PRLM-es, PRSVM-bp, PRSVM-es, iVECTOR
	PO	PRLM-es, PRSVM-bp, PRSVM-en, PRSVM-es, PRSVM-pt, iVECTOR
	EC	PRLM-bp, PRLM-en, PRLM-es, PRSVM-bp, PRSVM-es, iVECTOR
	EO	PRLM-en, PRLM-es, PRLM-pt, PRSVM-bp, PRSVM-es, iVECTOR

Table 2. Sub-systems selection in the *primary* and *contrastive1* submission for each evaluation condition. In the case of the *contrastive1* system, zt-norm is applied to the phonotactic sub-systems.

References

1. Rodríguez-Fuentes, L. J., Brümmer, N., Penagarikano, M., Varona, A., Diez, M., Bordel, G.: The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE). URL: `http://http://iberspeech2012.ii.uam.es/images/PDFs/albayzin_lre12_evalplan_v1.3_springer.pdf` (2012)
2. Rodríguez-Fuentes, L. J., et al.: Multi-site heterogeneous system fusions for the Albayzin 2010 language recognition evaluation. IEEE 2011 Automatic Speech Recognition and Understanding Workshop (ASRU) (2011)
3. Zissman, M.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing, vol. 4(1), pp. 31-44 (1996)
4. Li, H., Ma, B., Lee, C.-H.: A vector space modeling approach to spoken language identification. IEEE Transactions on ASLP, vol. 15, no. 1, pp. 271–284 (2007)
5. Martínez, D., Plchot, O., Burget, L., Glembek, O., Matejka, P.: Language Recognition in iVectors Space. in Proc. Interspeech 2011, Firenze, Italy (2011)
6. Rodríguez-Fuentes, L. J., et al.: The BLZ Submission to the NIST 2011 LRE: Data Collection, System Development and Performance. in Proc. Interspeech 2012, Portland, US (2012)
7. Meinedo, H., Neto, J.: Audio Segmentation, Classification and Clustering in a Broadcast News Task, in Proc. ICASSP 2003, Hong Kong (2003)
8. Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I., Neto, J: The L2F Broadcast News Speech Recognition System. in Proc. Fala2010, Vigo, Spain (2010)
9. Tong, R., Ma, B., Li, H., Chang, E. S.: Selecting phonotactic features for language recognition. in Proc. Interspeech 2010, pp. 737-740 (2010)
10. Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., Dumouchel, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. in Proc. Interspeech 2009, pp. 1559-1562 (2009)
11. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1435-1447 (2007)
12. Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., Dehak, R.: Language Recognition via Ivectors and Dimensionality Reduction, in Proc. Interspeech 2011, Firenze, Italy (2011)

13. Torres-Carrasquillo, P. A. et al.: Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features. in Proceedings of ICSLP 2002, pp. 89-92, Denver, Colorado (2002)
14. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing, 16(5), pp. 980-988 (2008)
15. Abad, A.: The L^2F Language Recognition System for NIST LRE 2011. In The 2011 NIST Language Recognition evaluation (LRE11) Workshop, Atlanta, US (2011)
16. Campbell, W. M., Sturim, D. E., Reynolds, D. A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, vol. 13(5), pp. 308-311 (2006)

Description of the MSMP-LaBRI system for albayzin 2012 lre

Jean-Luc Rouas

LabRI UMR 5800
351 Cours de la libération
Talence Cedex
France
rouas@labri.fr

1 PRSVM systems

Using the phonetic recognisers from the Speech Group at Brno University [1], we built three PRSVM systems. The phonetic recognisers used are for English, Hungarian and Russian.

The phonetic string from each of these decoders are used to build an 5gram "world" model using the SRI-LM toolkit [2]. The probabilities of occurrence of each of the 5grams are then computed on each file using the world 5gram model. These probabilities are then stacked into a sparse feature vector of huge dimension. These vectors are used to train language pair discriminant SVMs using the LibSVM library [3].

2 PLP-GSV SVM

PLP features are computed for each training file using the HTK toolkit http://htk.eng.cam.ac.uk/. They are then used to train a UBM-GMM using 512 gaussians. The adapted means of the UBM are computed on each training file, resulting in a 512-dimensional supervector per file. These supervectors are used as input features to compute a SVM for each language pair.

3 Prosodic model

The prosodic model is issued from [4]. Tokens describing F0 and energy variations as well as relative duration are computed on segments issued from the DFB algorithm [5]. These tokens are used to build a 5gram "world" model using data from all target languages. As for the PRVSM model, probabilities of occurrence of each of the 5grams are computed and used as a feature vector. The resulting vectors are used to train SVM models for each language pair.

4 Audio segmentation

Audio Segmentation is usually performed to automatically identify speech and music segments within an audio track. For the Albayzin 2012 LRE task, I used it to select data containing speech, with guesses on the kind of background.

The audio segmentation is based on Hidden Markov Models (HMM). 4 states HMM are used for this task, each state being modelled by a gaussian mixture with 256 components. The features used are classical speech processing features, Perceptual Linear Prediction coefficients (PLP). We used 12 coefficients together with their first and second-order derivatives.

4.1 Models training

The models were trained using a French broadcast corpus comprising a total of 50 hours designed for the ESTER evaluation campaign [XXX]. The available labels on this corpus are :

- acappella
- advertising
- applause
- jingle
- laugh
- multiple speech
- music
- other
- speech

Most labels are self-explanatory, multiple speech is used when several people speak at the same time, other is for sounds that do not pertain to any other category.

The system is somehow similar to a phonetic decoder, therefore it assumes that only one class may be encountered at a time. For this reason, we use mixed-event classes, i.e. music+speech, jingle+music, etc. in order to take into account simultaneous classes. As the number of classes significantly improves while considering mixed events, we had to select a short number of sufficiently represented mixed events (i.e. with more than 100 seconds of training data).

4.2 Recognition

During the recognition process, the test file is segmented and labelled using the viterbi algorithm. Then the labels corresponding to each of the target classes (e.g; speech and music) are merged from the output stream: if the target class is speech, it includes for instance speech+music, speech+laugh, speech+advertising, and so on. Afterwards, a post-processing scheme is applied in order to remove too short segments (i.e. under half a second of duration).

4.3 Evaluation

The audio segmentation system was tested on the development set of the same database. In the test set, there is a total of 19 hours of data, including almost 9 hours of speech and 12 hours of music (music and speech segments may overlap).

On this data, it achieved an F-measure of .928 (recall is 96.1 % and precision 89.6 %).

4.4 Use for albayzin LRE task

For the purpose of this evaluation, only the segments containing speech are of interest. The output of the system also gives us information on the kind of background. We have considered 4 event classes to train separate models :

- music+other+speech
- music+speech
- other+speech
- speech

Then, all the previously described models were trained for these different conditions.

The development and evaluation data was also segmented automatically using the audio segmentation algorithm with the same amount of event classes, resulting in 4 development and evaluation sets.

Thus, each system was also evaluated on each development (and evaluation) set. For instance, instead of having 3 PRSVM systems, I have 3 times 4 (training events classes) = 12 systems, tested on 4 development (or evaluatino) sets. This results in 12 times 4 = 48 outputs for the PRSVM systems.

5 Fusion

The linear fusion and conversion to llr are achieved using the FoCal toolkit [6]

As I also used the provided clean and noise labels to build all the systems, we have a total of :

- 3 PRSVM system outputs using no background knowledge
- 6 PRSVM system outputs using clean and noise labels
- 48 PRSVM system outputs using audio segmentation labels
- 3 PRSVM-lattices system outputs using no background knowledge
- 6 PRSVM-lattices system outputs using clean and noise labels
- 48 PRSVM-lattices system outputs using audio segmentation labels
- 6 PRLM system outputs using clean and noise labels
- 48 PRLM system outputs using audio segmentation labels
- 1 PLP-GSV system outputs using no background knowledge
- 2 PLP-GSV system outputs using clean and noise labels
- 16 PLP-GSV system outputs using audio segmentation labels
- 1 Prosodic system outputs using no background knowledge

– 2 Prosodic system outputs using clean and noise labels
– 16 Prosodic system outputs using audio segmentation labels

This results in a total of 206 outputs to be fused.

The primary system is the result of the fusion of these outputs.

Contrastive system 1 is the fusion of the 19 prosodic systems.

Contrastive system 2 is the fusion of all the systems build on the clean/noisy labels.

Contrastive system 3 is the fusion of all the prosodic systems build on the clean/noisy labels.

References

1. Matejka, P., Burget, L., Schwarz, P., Cernock, J.: Brno university of technology system for nist 2005 language recognition evaluation. In: Odyssey: The Speaker and Language Recognition Workshop. (2006) 57–64
2. Stolcke, A.: Srilm - an extensible language modeling toolkit. In: ICSLP. (2002) 901–904
3. Chang, C.C., Lin, C.J.: Libsvm : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3) (2011)
4. Rouas, J.L.: Automatic prosodic variations modelling for language and dialect discrimination. IEEE Transactions on Audio, Speech and Language Processing **15**(6) (2007) 1904–1911
5. Andr-Obrecht, R.: A new statistical approach for automatic speech segmentation. IEEE Transactions on Acoustics, Speech and Signal Processing **36**(1) (1988) 29–40
6. Brmmer, N., du Preez, J.A.: Application-independent evaluation of speaker detection. Computer Speech and Language **20** (2006) 230–275

The THU-EE Language Recognition Systems for Albayzin LRE 2012

Jia Liu, Liang He, Jiaming Xu,
Weiwei Liu, and Wei-Qiang Zhang

Department of Electronic Engineering,Tsinghua University,
Beijing 100084, China
{liuj,heliang,wqzhang}@tsinghua.edu.cn
liu-ww10@mails.tsinghua.edu.cn

Abstract. This paper describes the language recognition systems from the Department of Electronic Engineering, Tsinghua University (THU-EE) for Albayzin LRE 2012. Three systems are submitted for the evaluation, all of which are based on fusion of multiple subsystems. We describe each subsystem briefly and give their configurations. The processing speed of each system is also given in the paper.

Keywords: THU-EE, Albayzin LRE 2012

1 Introduction

This paper describes the language recognition systems from the Department of Electronic Engineering, Tsinghua University (THU-EE) for Albayzin 2012 Language Recognition Evaluation (LRE).

Our submissions are built on the following feature vectors:

- MFCC 7 + SDC 7-1-3-7: 7-dimensional MFCC is concatenated with shifted delta cepstrum (SDC) 7-1-3-7 [1], forming a 56-dimensional feature vector.
- TRAP: describes a segment of temporal evolution of critical band spectral densities within a single critical band [2].

Our submissions are built on the following classifiers:

- PPR-SVM: Parallel phone recognizer followed by support vector machine (PPR-SVM) [3].
- I-VEC: I-vector system [4] borrowed from speaker recognition.
- MCSK: Multiple coordinate sequence kernel (MCSK) system [5]. This system is similar to GSV-SVM, but using first- and second-order Baum-Welch statistics as supervectors.

2 Detailed subsystem descriptions

2.1 PPR-SVM

This subsystem is based on the phoneme recognition and SVM N-gram language modeling approach described in [6]. The Temporal Pattern Neural Network phone decoders developed by Brno University of Technology (BUT) for Czech(CZ), Hungarian (HU) and Russian (RU) is used as the frontend to compute phone lattice [7]. Then we use the maximum relevancy criterion based on mutual information to select the most discriminative N-gram patterns. We use N-grams up to 3 to perform language recognition, and weight the tri-gram probabilities using the term frequency log likelihood ratio method (TFLLR) as in [8]. We use the SVMTorch to train language models on the basis of the one-versus-rest strategy.

2.2 I-VEC

This subsystem is based on the i-vector [4] system which is popular and performing well in SRE current search. The used features consists of MFCC 7 + SDC 7-1-3-7. The used classifiers include both of cosine kernel scoring and SVM. Thanks to combination technology in i-vector level, we obtained the better performance than both of two, respectively.

2.3 MCSK

This subsystem has only one configurations, which uses MFCC 7 + SDC 7-1-3-7 feature vectors.

3 Fusion

We use a GMM + MMI backend [9] and FoCal tookit [10] for score calibration and fusion. To alleviate the unbalance of the development data, we use equalized linear discriminate analysis for dimension reduction and de-correlation. Then the GMM trained with MMI criterion is used as classifier.

4 Submission systems

The primary system selects ivector subsystem scores which use GMM + MMI backend and FoCal tookit for fusion. The contrast1 system consists of HU, RU frontend PPR-SVM, ivector and ivector-svm subsystem that apply GMM + MMI backend and FoCal tookit backend. The contrast2 system is similar to the contrast1 system, except that it is only based on GMM + MMI fusion strategy. The contrast3 system is fused with HU, RU frontend PPR-SVM and ivector subsystem, which seem more robust in the evaluation. The contrast4 system score is the fusion of all the subsystems' scores.

5 Processing time

Performances of all subsystems were measured separately on only one core of an Intel Core 2 Quad CPU 2.4GHz and 2 GB memory. Results are shown in Table 1. Note that the real time (RT) factor of the primary fusion system is less than the sum of all subsystems. It is because several subsystems share some common processing stages.

Table 1. *Per subsystem processing speed in real time factors.*

Subsystem	RT
PPR-SVM	0.1
I-VEC	0.05
MCSK	0.01

References

1. P. A. Torres-Carrasquillo, "Language identification using Gaussian mixture models," Ph.D. dissertation, Michigan State University, 2002.
2. Schwarz P, Matejka P, Cernocky J. Towards lower error rates in phoneme recognition. Proceedings of International Conference on Text, Speech and Dialogue, Brno, 2004. 465-472.
3. H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 271–284, Jan. 2007.
4. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788 –798, May. 2011.
5. L. He, Y. Deng, and J. Liu, "Method and system for speaker recognition based on multiple coordinate sequence kernel," China Patent, Sept. 1, 2009, CN101640043.
6. F. S. Richardson and W. M. Campbell, "Language recognition with discriminative keyword selection," in *Proc. ICASSP*, 2008, pp. 4145 – 4148.
7. P. Schwarz, Phoneme recognition based on long temporal context, Ph.D. dissertation, Faculty of Information Technology BUT, http://www.fit.vutbr.cz, Brno, CZ, 2008.
8. W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," in *Advances in Neural Information Processing System 16*, S. Thrun, L. Saul, and B. Scholkopf, Eds. MIT Press, 2004.
9. W.-Q. Zhang, T. Hou, and J. Liu, "Discriminative score fusion for language identification," *Chinese Journal of Electronics*, vol. 19, no. 1, pp. 124–128, Jan. 2010.
10. N. Brummer and D. van Leeuwen, On calibration of language recognition scores, in Proceedings of Odyssey - The Speaker and Language Recognition Workshop, 2006, pp. 1C8.

The UVigo-GTM System for the Albayzin'12 Language Recognition Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
Universidade de Vigo
E.E. Telecomunicación, 36310 Vigo (Spain)
{plopez,ldocio,carmen}@gts.uvigo.es

Abstract. This paper describes the UVigo-GTM system for the Albayzin 2012 Language Recognition Evaluation. It consists of the fusion of seven different subsystems, three phonotactic subsystems and four acoustic subsystems. The phonotactic ones are based on the Phone Recognition and Language Modeling (PRLM) approach, and the Brno University of Technology (BUT) phoneme recognizers developed for Hungarian, Czech and Russian languages are used. One of the acoustic subsystems computes the likelihood among a speech utterance and the acoustic models for each target language, which are trained by expectation-maximization and maximum mutual information. Another acoustic subsystem models the data by means of Gaussian mixture models and classifies the audio utterances using a support vector machine. The third acoustic subsystem is similar to the previous one, but the Fishervoices transformation is applied. The last acoustic subsystem is based on the iVector approach.

Keywords: Language recognition, Phone Recognition and Language Modeling, Fishervoices, iVectors.

1 Introduction

The Spoken Language Recognition (SLR) task has experimented a huge development in the last years. To compare the quality of the different SLR systems around the world, and to establish the baseline of performance capability for language recognition technology on conversational telephone speech, NIST has coordinated a language recognition evaluation every two years since 1996. The Spanish Thematic Network on Speech Technology (RTTH) coordinated similar language recognition evaluations in 2008 and 2010 called Albayzin LRE evaluations. The main difference with NIST is that the languages to be recognized are Spanish, Catalan, English, Basque, Galician and Portuguese, and they are extracted from multi-speaker TV broadcast recordings. The Albayzin 2012 Language Recognition Evaluation differs from the prior ones in that there exist a mismatch between the train and test acoustic conditions, leading to more difficulties. Specifically, the application domain moves from TV Broadcast speech to any kind of speech found in the Internet. Thus, audio signals for development

and evaluation were extracted from YouTube videos, which are heterogeneous regarding duration, number of speakers, environmental noise/music, channel conditions, etc. Besides speech, signals may contain music, noise and any kind of non-human sounds.

In this new evaluation different test conditions are proposed:

- Depending on the set of languages that are allowed to appear in the audio signal, two types of recognition tests are defined: open-set and closed-set recognition.
- Depending on the availability of training materials for target languages, two different conditions are defined: plenty of training and empty of training.

Combining these two tasks and two recognition modalities, four evaluation tracks are defined:

- Plenty of Training (6 target languages), Closed-Set Recognition (briefly, PC)
- Plenty of Training (6 target languages), Open-Set Recognition (briefly, PO)
- Empty Training (4 target languages), Closed-Set Recognition (briefly, EC)
- Empty Training (4 target languages), Open-Set Recognition (briefly, EO)

This paper presents the UVigo-GTM system for the Albayzin 2012 Language Recognition Evaluation developed for these four tasks. The system consists of the fusion of seven different subsystems, three phonotactic subsystems and four acoustic subsystems.

The rest of the paper is organized as follows: Section 2 describes the developed system, detailing each one of the seven individual subsystems and the scores fusion and calibration methods of these subsystems. Section 3 specifies the training and development data used. Section 4 presents the system performance on the development data. Section 5 indicates the computational cost of the system.

2 UVigo-GTM Systems

The UVigo-GTM language recognition systems are composed by seven subsystems: three PRLM subsystems with phoneme recognizers developed for Czech (PhonotacticCZ), Hungarian (PhonotacticHU) and Russian (PhonotacticRU); and four acoustic subsystems, one that computes the likelihood among an utterance and the acoustic models for each target language, which are trained by expectation-maximization and maximum mutual information (GMM-MMI-LH), one that classifies using a support vector machine where data is represented by supervectors obtained by adapting a Gaussian mixture model (GMM-SVM), one that follows the same approach as the previous one but applies the Fishervoices transformation to the data (Fishervoices-SVM), and one that applies the iVector approach (iVectors).

The output scores of the subsystems are calibrated and fused in order to obtain the final language recognition scores.

2.1 Phonotactic subsystems

Three phonotactic subsystems under a phone recognition followed by language modeling (PRLM) approach have been used. These subsystems use the Brno University of Technology (BUT) phoneme recognizer, which is based on ANN/HMM and temporal patterns (TRAPS) with split temporal context (STC) [11]. Specifically, the phone tokenization of the speech data is performed using the BUT recognizers trained for Czech, Hungarian and Russian languages, which are publicly available.

In the training stage, each phonetic tokenizer is used to extract the phonotactic features of each target language for each environmental condition (clean and noisy) in the "Plenty of traing" condition; in the "Empty of training" condition, there are no distinction between environmental conditions. The phonetic tokenizer outputs a phone lattice for each speech utterance, which is used to extract and model the phonotactic statistics of the target language. Next, in the "Plenty (Empty) of training" condition, the language phonotactics are modeled using a phone 4-gram (3-gram) back-off language model (LM) that is trained from phone lattices using an interpolated Witten-Bell discounting method. The SRILM toolkit has been used for that purpose [13]. A different 4-gram (3-gram) language model is trained for each target language and environmental condition, resulting in a total of 12 language models (6 languages x 2 conditions) in the case of the "Plenty of training" condition, and 4 language models in the case of the "Empty of training" condition. In the test stage, the phone lattice of a given test utterance is obtained and the likelihood of each target language model is evaluated.

2.2 Acoustic subsystems

A parameterization consisting of 7 static Mel-Frequency Cepstral Coefficients (MFCC) concatenated to 49 Shifted Delta Cepstral (SDC) coefficients computed under a 7-1-3-7 configuration was used in the acoustic subsystems. Cepstral Mean Normalization (CMN) and RelAtive SpecTrAl (RASTA) filtering were applied in MFCC computation in order to reduce the effects of intersession variability. This normalization was different in the GMM-MMI-LH acoustic subsystem. where it was replaced for Cepstral Mean and Variance Normalization (CMVN), RelAtive SpecTrAl (RASTA) and feature warping [10] using a sliding warping window set to 3s.

Non-speech regions do not typically contain language-specific information. Then, a speech activity detector (SAD) was used to remove silence and non-speech frames before training and evaluating the acoustic models. The SAD detector included in the SHOUT toolkit [6], which is freely available under a GNU license, was used. This detector uses a bootstrapping SAD component so it does not use any parameters that require tuning on in-domain training data.

After extracting the acoustic features, the training data from all target languages was used to train a 512-mixture Universal Background Model (UBM). Also, six (four) UBMs, one per target language in the "Plenty of training" condition, where also trained with this training data. For the "Empty of training"

condition, the UBMs of the four target languages were obtained by adapting the UBM to data of these languages.

A description of the four acoustic subsystems is given below.

2.2.1 GMM-MMI-LH subsystem. The UBMs of each target language, which were trained by expectation-maximization, are re-trained by maximum mutual information (MMI) [5]. Then, the likelihoods of each speech utterance with these six GMM-MMI models are computed. Six (four) scores, one per target language in the "Plenty (Empty) of training" condition, are obtained for each speech utterance. A backend is applied to these scores for calibration, and this is done by means of the FoCal multiclass library [4].

2.2.2 GMM-SVM subsystem. The UBM is adapted to the training data, obtaining a set of GMMs whose means are concatenated, resulting in a set of training supervectors. These supervectors are used to train a support vector machine (SVM) that classifies the test supervectors (obtained in the same way as the training supervectors) into the six (four) different target languages of the "Plenty (Empty) of training" condition. SVM training and prediction are performed using the library for large linear classification LIBLINEAR [3]. The logistic regression model is trained, in order to be able to obtain the probability of each class as output. The data used for training the SVM is taken from the development dataset, in order to train the SVM with speech utterances that are similar to the ones that will be classified. The SVM outputs the probability of a speech utterance belonging to each target language.

No backend is used in the Closed Set condition because the scores are already calibrated, but in the case of the Open Set condition a backend is applied in order to obtain the calibrated scores for the Out-of-set class.

2.2.3 Fishervoices-SVM subsystem. Fishervoices is a transformation that projects the data into a less-dimensional and more discriminative subspace [8]. In this system, the UBM is adapted to the training data obtaining a set of GMMs, and then these GMMs are transformed by means of the Fishervoice projection, obtaining a set of supervectors that are used to train an SVM. Two transformation matrices are also obtained. In the test stage, once the GMMs of the test data are obtained, they are projected using the transformation matrices obtained in the training stage, and the resulting supervectors are classified using the trained SVM. The Fishervoices transformation has two free parameters, and $e_{PCA} = 500$ and $e_{LDA} = 50$ ($e_{PCA} = 500$ and $e_{LDA} = 40$) were selected after tuning for the "Plenty (Empty) of training" condition. As before, the SVM outputs the probability of a speech utterance belonging to each target language.

As in the previous system, a backend is applied only in the Open Set condition.

2.2.4 iVector subsystem. The iVector approach is implemented as described in [9]. The total variability matrix was estimated on zero and first-order sufficient

statistics computed according to [2] using the ALIZE/LIA_RAL toolkit [1]. This matrix is used to extract the total variability factors (iVectors) of the speech utterances as described in [2]. Then, these iVectors are normalized to unit length. The distributions of iVectors for individual languages were modeled by Gaussian distributions with a single within-class full covariance matrix shared by all the languages. In the test stage, the log-likelihood for each Gaussian language model is obtained for a given test utterance.

A total variability subspace of dimension 400 (200) was used for iVector extraction in the "Plenty (Empty) of training" condition.

2.3 Backend and fusion

Backend and fusion parameters were optimized on the development set of the Albayzin 2012 LRE. It was found that applying a backend to the output scores of the subsystems, except on the GMM-SVM and Fishervoices-SVM system, improved performance. Therefore, a generative Gaussian backend was applied to the scores provided by these subsystems. Then, the resulting log-likelihood values were fused by applying linear logistic regression under a multiclass paradigm, obtaining a set of calibrated scores. In the case of the Open Set condition the backend was also applied to the GMM-SVM and Fishervoices-SVM systems, in order to calibrate the scores for the Out-of-set class.

The FoCal Multiclass toolkit was used to estimate and apply the backend and calibration/fusion models [4].

3 Train and Development Data

All the UBMs (the UBMs for each target language used in the GMM-MMI-LH subsystem in the "Plenty of training" condition and the world model used in the GMM-SVM and Fishervoices-SVM subsystems) were trained using the training data of KALAKA-3 database [7]. The UBMs for each target language used in the GMM-MMI-LH in the "Empty of training" condition, as there is not a sufficient quantity of training data to train these UBMs, were obtained by adapting the world model to the development data of each of the target languages. The BUT Czech, Hungarian and Russian phone recognizers were trained on the SpeechDat-E databases [12]. The backends and fusion parameters were trained using the development dataset, and this dataset was also used to train the SVMs and the transformation matrices of the Fishervoices transformation.

4 Preliminary results

Recognition results of each individual subsystem and the fusion of the seven subsystems on the development data are presented in tables 1 (PC), 2 (PO), 3 (EC) and 4 (EO). The fact that results are better for the "Empty of training" condition is noticeable: it is due to the training of the subsystems, which was done

using the development data, and the test was run on this data too. Nevertheless, the way the subsystems are trained in the "Plenty of training" condition makes it easier for the system to have a good performance in mismatch conditions, as they are not trained for concrete acoustic conditions.

Table 1. Recognition results on development set: Plenty of training, closed set

System	F_{act}	F_{dis}	F_{cal}
PhonotacticCZ	0.246	0.230	0.068
PhonotacticHU	0.174	0.161	0.085
PhonotacticRU	0.190	0.176	0.079
GMM-MMI-LH	0.299	0.294	0.015
GMM-SVM	0.127	0.123	0.036
Fishervoices-SVM	0.152	0.150	0.009
iVectors	0.153	0.141	0.082
Fusion	0.045	0.045	10^{-8}

Table 2. Recognition results on development set: Plenty of training, open set

System	F_{act}	F_{dis}	F_{cal}
PhonotacticCZ	0.277	0.267	0.036
PhonotacticHU	0.208	0.192	0.080
PhonotacticRU	0.231	0.214	0.080
GMM-MMI-LH	0.332	0.330	0.006
GMM-SVM	0.170	0.154	0.100
Fishervoices-SVM	0.218	0.199	0.094
iVectors	0.174	0.162	0.075
Fusion	0.047	0.045	0.031

Table 3. Recognition results on development set: Empty of training, closed set

System	F_{act}	F_{dis}	F_{cal}
PhonotacticCZ	0.112	0.095	0.169
PhonotacticHU	0.069	0.064	0.083
PhonotacticRU	0.085	0.082	0.048
GMM-MMI-LH	0.104	0.097	0.066
GMM-SVM	0.237	0.234	0.012
Fishervoices-SVM	0.251	0.234	0.074
iVectors	0.078	0.077	0.024
Fusion	0.02	0.02	$2.24{\cdot}10^{-9}$

Table 4. Recognition results on development set: Empty of training, open set

System	F_{act}	F_{dis}	F_{cal}
PhonotacticCZ	0.209	0.185	0.13
PhonotacticHU	0.182	0.146	0.247
PhonotacticRU	0.2	0.171	0.162
GMM-MMI-LH	0.215	0.209	0.031
GMM-SVM	0.365	0.356	0.026
Fishervoices-SVM	0.368	0.358	0.028
iVectors	0.225	0.223	0.009
Fusion	0.082	0.082	$3.56{\cdot}10^{-10}$

5 Processing Speed

Real time factor is approximately $0.4 \times$ RT. To compute the processing speed of the UVigo-GTM system, ten test utterances were recognized and its execution time was divided by the duration of those utterances. This processing speed was obtained after executing this language recognition algorithm on a server with an Intel Xeon E5620 @ 2.4 GHz and 32GB of memory.

Acknowledgements This work has been supported by the Galician Regional Government (CN2011/019, 2009/062), Spanish Government (TEC009-14094-C04-04 and FPI grant BES-2010-033358), and the European Regional Development Fund.

References

1. Bonastre, J.-F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N., Fauve, B. and Mason, J., ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition, In: Proceedings of Odyssey: the Speaker and Language Recognition Workshop (2008). `http://mistral.univ-avignon.fr/index_en.html`
2. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., Front-End Factor Analysis for Speaker Verification, IEEE Transactions on Audio, Speech, and Language Processing, vol. 19(4), 788-798, (2011).
3. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J., LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research, vol. 9, 1871–1874 (2008). Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`
4. FoCal, toolkit for evaluation, fusion and calibration of statistical pattern recognizers (2008), `http://sites.google.com/site/nikobrummer/focal`
5. Hubeika, V., Szöke, I., Burget, L. and Cernocký, J., Maximum Likelihood and Maximum Mutual Information Training in Gender and Age Recognition System, In: Proceedings of the 10th international conference on Text, speech and dialogue, 496–501 (2007).
6. Huijbregts, M., SHoUT Toolkit Web Page. `http://shout-toolkit.sourceforge.net`
7. KALAKA-3. Speech database created for the Albayzin 2012 Language Recognition Evaluation. Produced by the Software Technologies Working Group (GTTS), University of the Basque Country UPV/EHU, Spain.
8. Lopez-Otero, P., Docio-Fernandez, L. and Garcia-Mateo, C., A Fishervoice-SVM Language Identification System, Lecture Notes in Computer Science, vol. 7243/2012, 381–391 (2012).
9. Martinez, D., Plchot, O., Burget, L., Glembek, O. and Matejka, P., Language Recognition in iVectors Space, In: Proceedings of Interspeech, 861–864 (2011).
10. Pelecanos, J. and Sridharan, S., Feature Warping for Robust Speaker Verification, In: Proceedings of Odyssey, 213-218 (2001).
11. Schwarz, P., Phoneme Recognition Based on Long Temporal Context, Ph.D. Thesis, Brno University of Technology, 2009. `http://speech.fit.vutbr.cz/cs/software/phoneme-recognizerbased-long-temporal-context`

12. SpeechDat-E database, http://www.fee.vutbr.cz/SPEECHDAT-E/
13. Stolcke, A., SRILM – An Extensible Language Modeling Toolkit, In: Proceedings of International Conference on Spoken Language Processing, vol. 2, 901–904 (2002).

Albayzin 2012 LRE @ ViVoLab UZ

David Martínez, Eduardo Lleida, Alfonso Ortega

Voice Input Voice Output Laboratory (ViVoLab)
Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
(david | lleida | ortega)@unizar.es

Abstract. This paper describes the systems submitted to the Albayzin 2012 Language Recognition Evaluation. The main novelty of this evaluation is the inclusion of audio files coming from a real scenario such as internet. ViVoLab has built 8 systems of different nature: 4 are based on an acoustic approach, where the signal frequency information is modelled with the SDC, iVectors are extracted to reduce dimensionality, and these are finally used to train models for each language, being these models Gaussian distributions in one of the systems, and a Restricted Boltzmann Machine in the other three; 3 are based on a phonotactic approach, where the statistics of apparition of the phonemes recognised by a phoneme recogniser are used to train SVMs: the last system is based on prosody and formant information, where the pitch, energy, duration and formants are modelled with Legendre polynomials and these are used to train an iVector-based Gaussian classifier identical to the acoustic one. In most of the submissions a Gaussian backend is used as second-step classifier, and finally, scores are calibrated with MultiFocal toolkit.

Keywords: Language Recognition Evaluation, iVector, Phonotactic LRE, Restricted Boltzmann Machine, Prosodic LRE, Gaussian Backend

1 Introduction

State of the art language recognition systems achieve very high performance on clean speech. To foster new challenges to the community, Albayzin 2012 language recognition evaluation (LRE) proposes new tasks with higher difficulties [1]. The first difference with previous evaluations is that development and evaluation data are extracted from YouTube videos and thus the content is very heterogeneous, with a high variety of speakers, channels, noises, etc. The second difference is that there are two tasks differentiated by the training conditions. In the first one, named *Plenty of Training*, the target data are the languages spoken in the Iberian peninsula, that is, Basque, Catalan, English, Galician, Portuguese, and Spanish. For this task, a big amount of data is provided (around 18 h per language) for training, and also development data taken from YouTube (between 100-200 signals per language and 500 for Out-Of-Set languages (OOS)). In the second one, named *Empty Training*, the target languages are French, German, Greek and Italian. In this case no training data are provided to the participants,

and only a development dataset recorded from YouTube is available in these languages (between 100 and 200 recordings per language). The data from the first task can be also used. Note that the OOS signals are the same for both tasks, and that no other external data can be used in this evaluation, except in the case of using auxiliary systems like a phoneme recogniser. The database recorded for this evaluation is known as KALAKA-3 [2].

ViVoLab is the second time that takes part in the Albayzin LRE [3]. For Albayzin 2012 LRE, 8 systems have been built, fused in different ways to send a total of four submissions for the Plenty of Training-Closed Set (PC) condition, and two for each of the rest, that is, Plenty of Training-Open Set (PO), Empty Training-Closed Set (EC), and Empty Training-Open Set (EO). Our 8 systems are divided as follows. 4 of them are acoustic systems, where the frequency information of the signal is extracted via the shifted delta coefficients (SDC) [4]. The SDC are used to train an iVector system that allows to obtain fixed-length low-dimension vectors, the iVectors [5], that contain information about both the language and the channel. Simple classifiers can be built with the iVectors. In our case, one of our systems uses a Gaussian distribution for each language, and in the other three, Restricted Boltzmann Machines (RBM), trained as a self-content classifier in a generative fashion [6]. 3 of the 8 systems are phonotactic, where the frequency of apparition of the phonemes recognised by a phoneme recogniser are used to model each language. The phoneme recognisers are trained on Hungarian, Czech and Russian. Once we obtain the counts of the phonemes for each audio recording, a one-versus-all support vector machine (SVM) is trained for each language [7]. The last system is based on prosody and formant information. In this approach, pitch, energy, duration and formants F1 and F2 are modelled with Legendre polynomials, that are the input to a iVector-based system identical to the acoustic one [8].

RBMs have not been used historically in speech processing, but after the work in [9, 10], where they are employed successfully for phone recognition, the speech community has started to be interested in them. Especially, since the strong conditional independence and unrealistic assumption imposed by hidden Markov models (HMM), in which every observation is independent of the rest given the state that generated it, can be relaxed. In spite of being computationally demanding, the increasing popularity of parallel computing architectures like GPUs together with the relaxation of conditional independence make RBMs and their deep version, deep belief networks (DBN), a very promising approach for speech technologies. Also these ideas have started to be studied in other fields of speech processing, like speaker recognition [11, 12]. In this evaluation, a first contact of RBMs with language recognition is made.

The addition of formants F1 and F2 to the prosodic system is a new contribution that has shown to be successful in internal experiments carried out in our group. It is a continuation of the work conducted in [8], and we expect to publish soon the whole experiments once they are conclusive enough.

The primary submission of the PC condition consists of 5 systems, the acoustic with the Gaussian classifier for iVectors, the three phonotactic SVMs, and

the prosodic one, using a Gaussian backend as final step. The first contrastive includes the 8 systems commented in the previous paragraph, also with a Gaussian backend. The second and third contrastive systems are identical to the primary and first contrastive, respectively, but no Gaussian backend is used. For the PO, EC and EO conditions, only the primary and one contrastive submissions have been sent, and they are identical to the PC condition, also including a Gaussian backend previous to the fuser-calibrator.

In the following sections each of the systems will be described in detail.

2 System Descriptions

2.1 Acoustic Systems

The 4 acoustic systems share the iVector extractor procedure. First, standard 7 Mel Frequency Cepstral Coefficients (MFCC) (including C_0) are obtained. Vocal Tract Length Normalization (VTLN) and Cepstral Mean and Variance Normalization are applied in MFCC computation. Then, SDCs with usual 7-1-3-7 configuration are obtained, and concatenated to MFCCs, to obtain a final feature vector of 56 coefficients.

The next step is to train a UBM model to collect zero and first order statistics. 2048 components of a Gaussian mixture model (GMM) are trained using maximum likelihood (ML). Then, a 600-dimension iVector extractor is trained, and iVectors are extracted for each of the recordings in the database. Next the two different approaches for classification are explained.

Gaussian classifier A Gaussian distribution is trained for each of the target languages, with shared covariance matrix equal to the within class covariance matrix of the training data. This is the same system as used in [13].

Restricted Boltzmann Machine An RBM is a type of Boltzmann Machine [14] where there are no visible-visible or hidden-hidden conexions. It has been traditionally used to model probability distributions generatively, by minimizing an energy-based function, but it has been shown that they also can be trained directly for classification [6]. In this work three different Gaussian-Bernouilli RBMs [15, 16] have been generatively trained with a learning rate of 0.1, and a momentum of 0.5. The learning process proposed in [15] has been followed, where the energy function is given by

$$E(\mathbf{v},\mathbf{h}|\theta)=\sum_{i=1}^{N_v}\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{i=1}^{N_v}\sum_{j=1}^{N_h}W_{ij}h_j\frac{v_i}{\sigma_i^2}-\sum_{j=1}^{N_h}c_jh_j-\sum_{k=1}^{N_y}\sum_{j=1}^{N_h}U_{kj}h_jy_k-\sum_{k=1}^{N_y}d_ky_k, \tag{1}$$

where $\mathbf{v}$ is the vector of N_v visible units, $\mathbf{h}$ is the vector of N_h hidden units, $\mathbf{b}$ and $\mathbf{c}$ are the biases of the visible and hidden units, respectively, $\mathbf{W}$ is the matrix of weights connecting visible and hidden units, $\sigma_\mathbf{i}$ is the standard variance associated to visible unit i, $\mathbf{y}$ is the binary vector of N_y labels, $\mathbf{d}$ is the bias of

the labels, and **U** is the matrix of weights connecting label and hidden units. It is important to note that the visible-hidden connexions form a Gaussian-Bernouilli RBM and the label-hidden connexions form a Bernouilli-Bernouilli RBM.

The first RBM has been trained on 500 epochs, has 100 hidden neurons and the Gaussian noise variances have been kept fixed and equal to unity. The second and the third ones have been trained on 10000 epochs and the variances have been also updated, but whereas in the second the number of hidden neurons is 75, in the third it is 80. The reason to train different RBMs is empirical, after analysing which systems were performing better in the part of the development set used for testing. A practical guide of how to train an RBM and the influence of each of these parameters can be found in [17]. In figure 1 the structure of an RBM trained for classification can be seen.

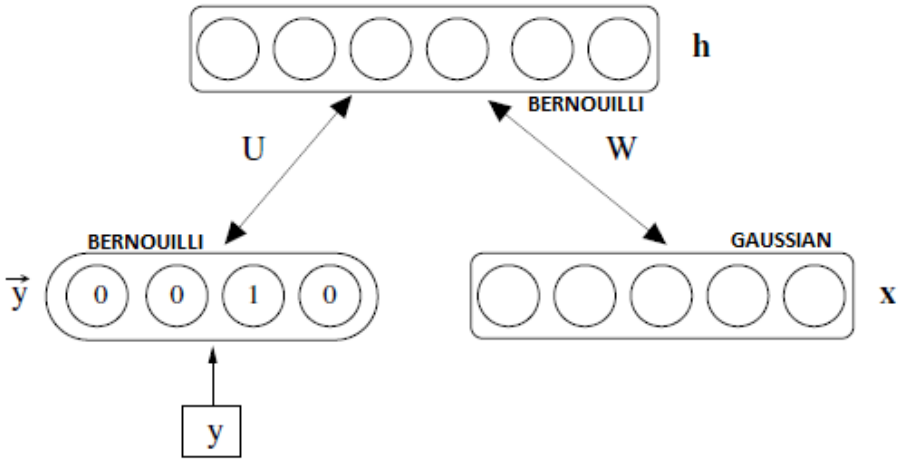


Fig. 1. Architecture of an RBM for classification. **x** are the visible units, **h** are the hidden units, **y** are the labels, **W** is the matrix with the weights for visible units, and **U** is the matrix with the weights for the labels. Note that **y** is a binary vector with the position of the true class equal to 1

2.2 Phonotactic Systems

Brno phoneme recogniser [18] trained on Hungarian, Czech and Russian has been used to obtain the sequence of phonemes of every file. Following the proposal of [7], the 40000 most frequent 3-grams in the training dataset are kept. Then, a linear SVM is trained with their counts, and in the evaluation phase, only these 40000 3-grams are considered. L2 regularization is applied with $C = 0.01$.

2.3 Prosodic and Formants System

This is a similar system to the one presented in [8], where pitch, energy and duration are extracted and modelled with Legendre polynomials, but with the main difference that in the present work formants F1 and F2 have been added

to the set of features. A similar experiment can be found in [19] for speaker verification. F1 and F2 are known as vocalic formants and the assumption made here is that their evolution in time can be a good cue for language discriminability. More experiments are being run at the moment but we can advance that higer formants give no further gains, probably because their calculation is noisier and also because F4 and higher formants are considered to be related with the speaker's vocal tract characteristics.

The order of Legendre polynomials is 5, what gives 6 coefficients per feature, except for duration which is just a scalar whose value is the number of voiced frames in the region of computation of the polynomial. Hence a final feature vector of 25 dimensions is obtained. Polynomials are obtained in fixed regions of 200 ms, every 10 ms. Only voiced frames are used for the calculation. Once all the features are extracted, the rest of the system is identical to the acoustic one based on iVectors with a Gaussian model for classification.

Another substantial difference with the work in [8], where Snack software was used to obtain pitch and energy, is that in this work a pitch and formant extractor built in our group is used. The pitch extraction procedure is based on average magnitude difference function (AMDF), and formants are calculated by finding the roots of the linear predicitve coding (LPC) polynomial, but in both cases a beam-search is used to find the best sequences of pitch and formants given a set of frquency candidates, applying a cost function based on local and global evidences. The main advantage of this beam-search is that a trajectory function that takes into account several frames can be incorporated into this cost function. No publication is availabe for the pitch extraction algrithm yet, and formant extraction procedure is detailed in [20].

2.4 Gaussian Backend, Fusion and Calibration

The Gaussian backend can be considered as a second classifier in cascade with the previous sytems, trained on the development dataset. In most of the current systems it has been found to be beneficial. Probably, in some cases it could be due to having development data available for training the Gaussian backend more similar to the test data than the training data. But, could the same results have been obtained if those data were used to train the first stage of the systems directly? In addition, development datasets are normally more balanced among languages than the training datasets, what can cause this second classifier to compensate for this fact. In this evaluation all classes are well balanced in the training dataset and in our system development process no big differences between systems with or without the Gaussian backend have been found. For this reason, in the PC condition, we have submitted each of the final systems twice, one with Gaussian backend and one without. In the cases where the Gaussian backend has been included, we have used MultiFocal toolkit [21].

For fusion and calibration, also MultiFocal toolkit has been used. In MultiFocal toolkit, the metric to be optimized is C_{llr}, and it is optimized via a discriminative logistic regression [21]. Also the Bilbao toolkit, given with the evaluation data, was tested, but no significant differences with MultiFocal were

found. It is worth to note that in this evaluation, unlike the NIST LRE evaluations, the priors are flat over all classes.

3 Submissions and Data

3.1 Submissions

Plenty-Closed

- Primary: fusion of 5 subsystems: acoustic with Gaussian classifier + SVM phonotactics in Hungarian, Czech, and Russian + prosodic and formants + Guassian backend.
- Contrastive 1: fusion of 8 subsystems: acoustic with Gaussian classifier + acoustic RBMs with 75, 80 and 100 hidden neurons + SVM phonotactics in Hungarian, Czech, and Russian + prosodic and formants + Guassian backend.
- Contrastive 2: same as primary but without Gaussian backend.
- Contrastive 3: same as contrastive 1 but without Gaussian backend.

Plenty-Open, Empty-Closed, and Empty-Open For these conditions only the primary and contrastive 1 submissions of PC condition were submitted.

3.2 Data

In all cases models for the target languages of the Plenty condition were trained. Only the KALAKA-3 training data were used for training. The OOS model and the models for target languages of the Empty conditions were trained only in the backend with the development data. In any case, development data were only used to train the backend. See figures 2 and 3 for more detail.

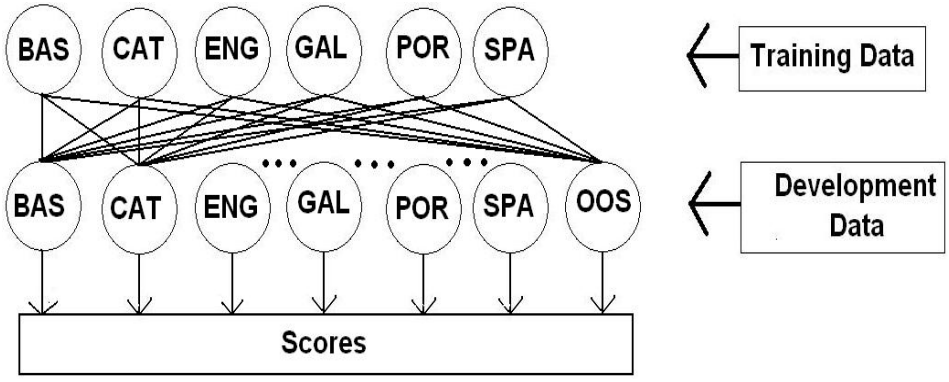


Fig. 2. Data distribution in the Plenty conditions

For selecting the systems, we divided the development dataset in two parts. One was used to train the Gaussian backend, fuse, and calibrate, and the other

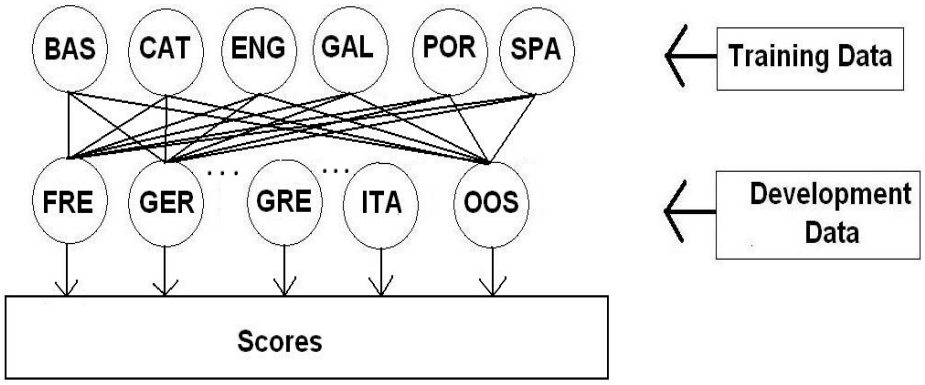


Fig. 3. Data distribution in the Empty conditions

was used to test. The selection was made empirically and the fusions that performed the best in our test set were chosen. For testing the evaluation data, we used all development dataset to train the Gaussian backend, fuse, and calibrate.

4 Processing Speed

I3A high-performance computing cluster has been used for all the experiments. This cluster is a group of

- 55 Dual Nehalem processors at 2.33 Ghz with 24 Gb RAM
- 44 Dual Xeon processors at 2.8 Ghz with 2 Gb RAM
- 11 Dual processors Dual Core Intel Woodcrest at 2.33 Ghz with 4 Gb RAM
- 12 Quad Core Intel Woodcrest processors at 3.0 Ghz with 8 Gb RAM
- 4 Dual processors Quad Core Intel Woodcrest at 2.66 Ghz with 16 Gb RAM
- 48 Dual processors 8-Core AMD Magny-Cours 6128 at 2.0 Ghz with 32 Gb RAM
- 2 Quad Core Intel Kentsfield at 2.33 Ghz with 8 Gb RAM
- 2 AMD Quad Opteron at 2.2 Ghz with 16 Gb RAM
- 2 12-core AMD Magny-Cours 6174 processors at 2.1 Ghz with 96 Gb RAM

In total, 1534 processors are available summing 3.5 Tb RAM. Condor software is used for parallelization.

Processing times have been calculated with a short list of files considering the minimum duration indicated in the evaluation plan, i.e. 30 s. This calculation has been made in a Dual Nehalem processor at 2.33 Ghz with 24 Gb of RAM. The real time (RT) factor breakdown for the individual stages of the systems is summarized in table 1. The breakdown for each subsystem is in table 2.

Considering that each of the systems is run in parallel, the total RT of our system will be given by the slowest one, that is, the prosodic and formants system, that sums a total of 2.9589xRT with the iVector classifier. Adding the fusion time, the total RT of our submitted systems is 2.9685xRT.

Task	xRT
SDC Extraction	0.2777
Pitch, Energy and Formants Computation	0.3543
Legendre Polynomials Computation	2.4033
Stats Computation	0.0748
iVector Extraction	0.1143
iVector Gaussian Classification	0.0122
iVector RBM Classification	0.0012
Lattice Generation	0.5271
SVM Classification	0.0016
Fusion	0.0096

Table 1. *RT factor breakdown into the individual stages of the system*

Subsystem	xRT
Gaussian iVector	0.4790
RBM iVector	0.4680
Phonotactic	0.5287
Prosodic and Formants	2.9589
Fusion	0.0096

Table 2. *RT factor for each individual subsystem*

5 Conclusions

This is the second participation of ViVoLab in the Abayzin LRE evaluations. This year, the aim is to work with signals coming from a real scenario like internet. In the Plenty condition, training data are available for all target languages. 8 systems have been built, our primary submission consists of 5 of them, and our contrastive submission consists of all of them. We have sent two more submissions identical to the previous ones but without backend. For the Empty condition, where only development data are available for the target languages, we have submitted the same systems but in all cases with a Gaussian backend, since this backend is the model trained on the target languages. As major novelties we have introduced RBM classifiers for acoustic iVectors and prosodic and formant information also modelled with iVectors.

6 Acknowldgements

This work was funded by the Spanish Ministry of Science and Innovation under project TIN2011-28169-C05-02.

References

1. L. J. Rodríguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, M. Díez, and G. Bordel, "The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE)", *Iberspeech, Madrid*, 2012.
2. KALAKA-3. Speech database created for the Albayzin 2012 Language Recognition Evaluation. Produced by the Software Technologies Working Group (GTTS), University of the Basque Country UPV/EHU, Spain.

3. D. Martínez, J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "I3A Language Recognition System for Albayzin 2010 LRE", *Interspeech, Florence*, 2011.
4. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", *International Conferences on Spoken Language Processing, Denver*, 2002.
5. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788-798, May 2011.
6. H. Larochelle and Y. Bengio, "Classification Using Discriminative Restricted Boltzmann Machines", *International Conference on Machine Learning, Helsinki*, 2008.
7. T. Mikolov, O. Plchot, O. Glembek, P. Matějka, L. Burget, J. Černocký, "PCA-based feature extraction for phonotactic language recognition", *Odyssey - The Speaker and Language Recognition Workshop, Brno*, 2010.
8. D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-Based Prosodic System for Language Identification", *ICASSP, Kyoto*, 2012.
9. A. Mohamed, G. Dahl, and G. Hinton, "Deep Belief Networks for Phone Recognition", *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, Whistler*, 2009.
10. A. Mohamed and G. Hinton, "Phone Recognition Using Restricted Boltzmann Machines", *ICASSP, Dallas*, 2010.
11. M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak, and P. Dumouchel, "First Attempt of Boltzmann Machines for Speaker Verification", *Odyssey - The Speaker and Language Recognition Workshop, Singapore*, 2012.
12. T. Stafylakis, P. Kenny, M. Senoussaoui, P. Dumouchel, "Preliminary Investigation of Boltzmann Machine Classifiers for SPeaker Recognition", *Odyssey - The Speaker and Language Recognition Workshop, Singapore*, 2012.
13. D. Martínez, O. Plchot, L. Burget, O. Glembek, P. Matějka, "Language Identification in iVectors Space", *Interspeech, Florence*, 2011.
14. G. Hinton, T. Sejnowski, and D. Ackley, "Boltzmann Machines: Constraint Satisfaction Networks that Learn", Technical Report CMU-CS-84-119, May 1984.
15. K. Cho, A. Ilin, T. Raiko, "Improved Learning of Gaussian-Bernouilli Restricted Boltzmann Machines", *International Conference on Artificial Neural Networks, Espoo*, 2011.
16. A. Krizhevsky, "Learning Multiple Layers Of Features From Tiny Images", Master Thesis, University of Toronto, 2009.
17. G. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines. Version 1", Technical Report UTML TR 2010-003, August 2011.
18. P. Schwarz, "Phoneme Recognition Based on Long Temporal Context", Ph.D. Thesis, Brno University of Technology, 2009.
19. N. Dehak, P. Kenny, and P. Dumouchel, "Continuous Prosodic Features and Formant Modeling with Joint Factor Analysis for Speaker Verication", *Interspeech, Antwerp*, 2007.
20. J. E. García, D. Ribas, A. Miguel, E. Lleida, and J. R. Calvo, "Beam-Search Formant Tracking Algorithm Based on Trajectory Functions for Continuous Speech", *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Buenos Aires*, 2012.
21. N. Brümmer, "FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual", http://sites.google.com/site/nikobrummer/focalmulticlass.

Albayzin Audio Segmentation Evaluation

System Descriptions

Audio Segmentation System by Aholab for Albayzin 2012 Evaluation Campaign

David Tavarez, Eva Navas, Daniel Erro, and Ibon Saratxaga

Aholab - Dept. of Electronics and Telecommunications. Faculty of Engineering.
University of the Basque Country. Alda. Urquijo s/n 48013 Bilbao.
{david,eva,derro,ibon}@aholab.ehu.es

Abstract. *This paper describes the system developed by Aholab Signal Processing Laboratory for the Albayzin 2012 audio segmentation evaluation campaign. A model based audio segmentation system that considers seven models (clean speech, narrow band speech, speech+music, speech+noise, music, silence, silence+music) has been built. This main system has a 21% of segmentation error in the development recordings. A post-processing step that analyzes the speech segments has been added to the main system, with a 27% improvement in the results of the music class, reducing the final segmentation error to 20%.*

Keywords: Automatic Audio Segmentation, Albayzin Evaluation Campaigns, Broadcast Speech

1 Introduction

Automatic audio segmentation is the process of dividing the audio stream into homogeneous sections according to its content. Depending on the application, the goal of the automatic segmentation process may be very diverse: separating speech from noise and music [1], separating male voice from female voice [2], separating the segments corresponding to different speakers [3], etc. Automatic audio segmentation has many applications and usually is used as a first preprocessing step to improve the performance of other systems like automatic speech recognition [4], speaker identification [5], content-based audio indexing and information retrieval [6],...

Three segmentation techniques have been mainly applied in automatic audio segmentation:

- *Metric based segmentation*: in these methods an acoustic distance measure is used to evaluate the similarity between two adjacent windows shifted along the audio stream. The audio stream is segmented at maxima of the distances and different regions are labeled in the stream accordingly. The most common distance measures applied are Kullback-Leibler Distance [7], Hotlling"s T^2 Distance [8] and Generalized Likelihood Ratio [9].
- *Model based segmentation*: a set of statistical models are built for each of the acoustic classes to be labeled (speech, noise, music, speech+noise,...).

Segmentation boundaries are assumed where there is a change in the acoustic class. Gaussian mixture models (GMM) [10], Hidden Markov Models (HMM) [11] and Artificial Neural Networks (ANN) [12] have all been applied to the model based audio segmentation task.

– *Model selection based segmentation*: these techniques apply a criterion to select the most suitable way to model an audio segment between using a single model for all the segment or using two different models. Usually the Bayesian Information Criterion (BIC) is applied for the model selection [13] [14], although some other hybrid criteria have also been proposed [15].

The Spanish Thematic Network for Speech Technologies (Red Temática en Tecnologías del Habla[1]) organizes each two years evaluation campaigns to objectively assess the validity of the algorithms developed for different speech technology related systems. In these campaigns, different research groups test their algorithms with a shared database, which allows for performance comparison and helps identifying new trends. Evaluations of text to speech systems [16] [17], automatic machine translation, language identification, diarization systems [18] and automatic audio segmentation [19] have been organized in past editions. This year one of the campaigns organized deals again with automatic audio segmentation. This paper describes the system developed by Aholab Signal Processing Laboratory to take part in this evaluation campaign.

The rest of the paper is organized as follows: section 2 presents the description of the system developed and the database used for training. The results obtained by the system are detailed in section 3 and finally section 4 deals with the conclusions of the work.

2 Description of the Proposed System

Figure 1 shows a diagram of the proposed Audio Segmentation System. In the first step a Viterbi segmentation is performed to locate the boundaries of the different acoustic events (speech, music, speech+music, specch+noise...) in the audio stream. Then, the labels obtained in the first step are treated in order to refine the boundaries, discard short silences and correct other minor mistakes. Finally, the speech segments are post-processed individually in order to identify those with low level music in the background.

In the following sections, the database used will be presented and each step of the system will be described in detail.

2.1 Database

The Albayzin 2012 database includes broadcast speech from the archive of Corporación Aragonesa de Radio y Televisión (CARTV). It contains around four hours of audio for development and another sixteen hours for testing. The Albayzin 2012 organization also provided for training the Catalan broadcast news

[1] http://www.rthabla.es/.

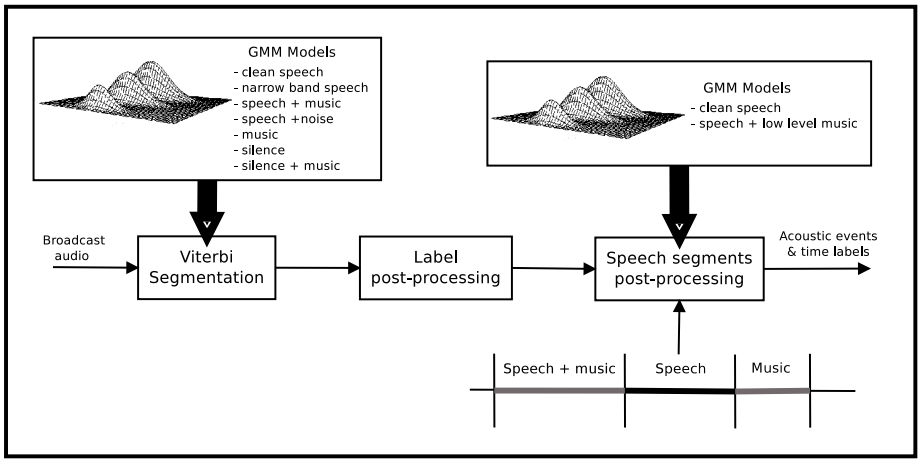


Fig. 1. Diagram of the proposed Audio Segmentation System

database from the 3/24 TV channel used in the 2010 Albayzin Audio Segmentation Evaluation [20]. The characteristics of this database are very different from the ones of CARTV. Due to the fact that the test database is also extracted from Aragón Radio database, only the CARTV database has been used in the development of the proposed audio segmentation system.

2.2 Viterbi Segmentation

A separate GMM model with 1024 mixtures was trained for clean speech, music, speech+noise, speech+music, narrow band speech, silence, and silence+music, using around four minutes of audio (for each model) of the development recordings of the database. The narrow band speech, silence and silence+music models have been included in order to improve the results of the system with these training recordings. The audio segments used to train the narrow band speech correspond to low pass filtered speech which is mostly labeled as speech+noise in the Albayzin 2012 reference labels. The silence and silence+music models have been trained with audio segments that correspond to pauses between sentences from the development recordings, taking into account the music in the background. These seven models are used in a Viterbi segmentation in order to detect the audio segments which contain the described acoustic events. Development experiments showed that the addition of derivatives of MFCC provided better segmentation results and that the use of 6 MFFC improved the location of the audio segments boundaries, therefore 6 MFCC with first and second derivatives were used for the classification. Finally, the segmentation labels obtained are mapped to the output classes that must be considered by the system (speech, music and noise) and are provided to the next module of the system, the label post-processing module.

2.3 Label post-processing

The aim of this step is to refine the boundaries, to discard short silences and to correct some minor errors made by the segmentation process. In the first place, silences shorter than 800 ms. are removed. These segments often correspond to pauses between words instead of boundaries between different audio segments, therefore they are eliminated and assimilated to the adjacent segments.

Development experiments also showed that short segments labeled as speech with noise in the segmentation process usually corresponded in fact to segments of speech with music in the background. This erroneous labeling occurred when the music was different from the one present in the training material of the model. In consequence, speech+noise segments shorter than 2.5 s. are relabeled in this post-processing step as speech+music. Similarly, segments labeled as silence+music longer than 800 ms. are relabeled as music. Finally, consecutive segments with the same label are unified in order to improve the performance of the post-processing step.

2.4 Speech segments post-processing

In this final step, the speech segments are processed in order to search for low level music in the background. When this low level music appears, the previous steps often mark the segments as clean speech, so a post-processing step as the one proposed in [21] is necessary to reduce the final error. First, a separate GMM G_{voice} with 1024 mixtures is trained using around fifteen minutes of real clean speech from the development recordings of the database. Similarly, a GMM G_{music} is trained using another fifteen minutes of speech with low level music. Finally, every segment labeled as clean speech is processed and if they are better modeled by G_{music} than by G_{voice} , the music label is added to the already existing speech label. Development experiments showed that, the use of 12 MFFC provided better results in this task, therefore, 12 MFCC with first and second derivatives were used for the classification in this step.

3 Results

This section presents the results obtained by the system in the development recordings of the Albayzin 2012 database. Table 1 shows the error of the proposed system after the segmentation and the label post-processing steps (main system) and after the speech segments post-processing module (post-processing). These results were obtained with the evaluation script provided by the Albayzin 2012 organization.

Table 2 presents the average CPU time required in order to process the development recordings. The time required for the post-processing step is also shown. These measures were made on a quad-core Intel Xeon 2.27 GHz computer with 32 GB memory.

As displayed in Table 1, there is a large variability in the obtained values, particularly for the main system, due principally to errors in the music and

Session	Main system	Post-processing	Session	Main system	Post-processing
1	25.67%	25.67%	17	1.61%	1.61%
2	39.68%	12.71%	18	5.02%	5.02%
3	43.99%	43.99%	19	28.36%	28.39%
4	20.74%	20.74%	20	24.71%	24.60%
5	28.29%	28.29%	21	28.36%	28.57%
6	35.82%	8.95%	22	2.24%	2.24%
7	37.09%	37.09%	23	8.92%	8.92%
8	3.25%	3.25%	24	11.85%	15.04%
9	36.25%	31.12%	25	31.73%	33.93%
10	34.44%	12.78%	26	8.75%	8.75%
11	18.65%	19.66%	27	30.55%	54.05%
12	37.84%	34.97%	28	13.51%	12.81%
13	28.54%	24.82%	29	31.32%	43.29%
14	6.74%	6.74%	30	9.68%	9.91%
15	19.60%	25.60%	31	16.30%	16.30%
16	8.61%	8.57%	32	40.09%	25.23%
ALL	21.45%	20.99%			

Table 1. Class Error Time of the system for the development sessions

Database	Main system	Post-processing
5 hours 16 minutes 34 seconds	2 hours 53 minutes 16 seconds	33 minutes 30 seconds

Table 2. CPU time required in order to process the development part of the database

noise labeling. By applying the speech segments post-processing the total error is reduced by 2.15% for the development part of the database, which barely proves the validity this step. However, if the acoustic events are evaluated individually, as it is displayed in tables 3 and 4, it becomes clearer how the post-processing step is decreasing the error. Table 3 shows the error of the proposed system after the segmentation step with the label post-processing and table 4 shows the error after the speech segments post-processing.

Error	Speech	Music	Noise
Missed Class Time	4.1%	38.7%	34.5%
False Alarm Class Time	1.1%	4.1%	9.9%
Total Class Error Time	5.19%	42.76%	44.33%

Table 3. Results of the main system for each speech, music and noise independently

As displayed in Table 3, the main source of the segmentation error resides on the music and noise detection (with 42.76% and 44.33% Class Error Time

respectively), while the speech labeling obtains good results (5.19%). In the post-processing step, low level music in the background of speech segments is located, and the result obtained can be seen in table 4.

Error	Speech	Music	Noise
Missed Class Time	4.1%	26.7%	34.5%
False Alarm Class Time	1.1%	4.3%	9.9%
Total Class Error Time	5.19%	31.02%	44.33%

Table 4. Results after post-processing for each speech, music and noise independently

As it can be seen in Table 4, the post-processing step decreases considerably the music Missed Class Time, reducing the obtained error by 27%, which proves the validity of this step for the enhancement of the labeling for music class.

4 Conclusions

This paper describes the system developed by Aholab Signal Processing Laboratory to take part in the Albayzin 2012 audio segmentation evaluation campaign. A model based audio segmentation system that considers seven models (clean speech, narrow band speech, speech+music, speech+noise, music, silence, silence+music) has been built. The error made by this system when working with the development recordings of the database from AragonRadio provided by Albayzin 2012 organization has been presented and analyzed. A main audio segmentation system has been built, with a 21.45% of Total Class Error Time in the development recordings. A post-processing step that analyzes the speech segments and tries to improve the labeling for music has been added to the main system, with a 27% improvement in the results for the music class, which translates in a 2.15% of Total Class Error Time Reduction in the whole system. With a 5.19% of speech class error, the proposed system presents as a good alternative to speech extraction in Broadcast Audio.

5 Acknowledgments

This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), and the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02).

References

1. Lu, L., Zhang, H., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Transactions on Speech and Audio Processing, 10(7), pp. 504 - 516. (2002)
2. Ore, B.M., Slyh, R.E., Hansen, E.G.: Speaker Segmentation and Clustering using Gender Information. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, pp. 1 – 8. San Juan, Puerto Rico. (2006)
3. Moattar, M. H., Homayounpour, M. M.: A review on speaker diarization systems and approaches. Speech Communication, 54(10), pp. 1065 - 1103. (2012)
4. Rybach, D., Gollan, C.: Audio segmentation for speech recognition using segment features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 4197 - 4200. Taipei, Taiwan (2009)
5. Reynolds, D. A., Torres-Carrasquillo, P.: Approaches and applications of audio diarization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 953 - 956. Philadelphia, USA. (2005)
6. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), vol. 2, pp. 5 - 8. Hong-Kong, China. (2003)
7. Siegler, M., Jain, U., Raj, B., Stern, R.: Automatic segmentation, classification and clustering of broadcast news audio, In: DARPA Speech Recognition workshop, pp. 97 - 99. Cantilly, USA. (1997)
8. Zhou, B., Hansen, J.: Efficient audio stream segmentation via the combined T^2 statistic and Bayesian information criterion. IEEE Transactions on Speech and Audio Processing, 13(4), 467 - 474. (2005)
9. Meignier, S. Moraru, D. Fredouille, C., Bonastre, J. F., Besacier, L.: Step-by-step and integrated approaches in broadcast news speaker diarization, Computer Speech and Language, vol. 20, pp. 303 - 330. (2006)
10. Aronowitz, H.: Segmental modeling for audio segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 393 - 396. Honolulu, USA. (2007)
11. Ramabhadran, B., Huang, J., Chaudhari, U., Iyenfgar, G., Nock, H. J.: Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives. In Interspeech, pp. 2589 - 2592. Geneva, Switzerland. (2003)
12. Meinedo, H., Neto, J.: A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models. In: Interspeech, pp. 237 - 240. Lisbon, Portugal. (2005)
13. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In: DARPA Broadcast News Transcription and Understanding Workshop, pp. 127 - 132. Lansdowne, USA. (1998)
14. Cettolo, M., Vescovi, M., Rizzi, R.: Evaluation of BIC-based algorithms for audio segmentation. Computer Speech and Language, 19(2), pp. 147 - 170. (2005).
15. Cheng, S., Wang, H.: METRIC-SEQDAC: A hybrid approach for audio segmentation. International Conference on Spoken Language Processing, pp. 1617 - 1620). Jeju Island, Korea. (2004)
16. Sainz, I., Navas, E., Hernez, I., Bonafonte, A., Campillo, F.: TTS Evaluation Campaign with a Common Spanish Database. In: Seventh International Language Resources and Evaluation (LREC'10), paper 454, pp. 2155-2160, Valletta, Malta. (2010)

17. Campillo-daz, F., Mendez, F., Arza, M., Docio, L., Bonafonte, A., Navas, E., Sainz, I.: Albayzin 2010 : a Spanish text to speech evaluation. In: 12th Annual Conference of the International Speech Communication Association (INTERSPEECH'11), pp. 2161 - 2164, Firenze, Italy. (2011)
18. Zelenák, M., Schulz, H., Hernando, J.: Albayzin 2010 Evaluation Campaign : Speaker Diarization. In: FALA 2010, pp. 301 - 304, Vigo (Spain). (2010)
19. Butko, T., Nadeu, C., Schulz, H.: Albayzin-2010 Audio Segmentation Evaluation Evaluation Setup and Results. In: FALA 2010, pp. 305 - 308, Vigo (Spain). (2010)
20. Butko, T., Nadeu, C.: Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing, 2011(1), pp. 1 - 10. (2011)
21. Tavarez, D., Navas, E., Erro,D., Saratxaga,I.: Strategies to Improve a Speaker Diarisation Tool. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4117-4121, Istambul (2012)

Albayzín Evaluations 2012. CAIAC-UAB Audio Segmentation System

Héctor Delgado and Javier Serrano

Center for Ambient Intelligence and Accessibility of Catalonia,
Campus UAB, 08193 Bellaterra, Spain
{hector.delgado,javier.serrano}@uab.cat
http://centresderecerca.uab.cat/caiac

Abstract. This paper describes the audio segmentation system developed by the Center for Ambient Intelligence and Accessibility of Catalonia (CAIAC), at the Autonomous University of Barcelona (UAB), for the Albayzin 2012 Audio Segmentation Evaluation. The evaluation task consists of segmenting spoken audio documents into three different acoustic classes: speech, background noise, and music, taking into account that more than one class may be present at a given time instant. The proposed system uses Gaussian Mixture Models (GMM) for modeling the audio classes. The overlapped regions are determined by means of a set of thresholds values tuned for this particular task. The system performance has been measured in terms of overall segmentation error, which reaches a 37.04 %

Keywords: Audio segmentation, GMM, MFCC

1 Introduction

Audio segmentation plays an important roll in some speech processing applications, such as Spoken Document Retrieval (SDR) and Automatic Speech Recognition (ASR) for broadcast shows. A suitable detection of audio classes and an accurate labeling will facilitate the subsequent systems operation. For instance, in automatic speech-to-text transcription, detecting different kinds of speech (e.g. clean speech or speech over background noise) will allow to use specialized acoustic models for those particular conditions, which will result in performance increases. In addition, audio regions identified as different classes than speech can be discarded previous to the recognizing stage. Therefore, accurate audio segmentation systems are needed to apply ASR and SDR in real-world environments.

This paper presents the CAIAC-UAB system for the 2012 Albayzín evaluation on Audio Segmentation. The developed system uses the standard MFCC feature extraction method, as well as GMMs for acoustic classes modeling. To deal with overlapping regions, some thresholds over log-likelihoods are taken into account for decision making.

The paper is organized as follows: section 2 provides details about the audio segmentation evaluation plan of the Albayzin 2012 evaluations. Section 3 describes the submitted system. The obtained results and a discussion are presented in section 4. Finally, section 5 gives the obtained conclusions.

2 Audio segmentation evaluation

This section describes the audio segmentation evaluation, as well as the audio databases and the scoring method used (consult [2] for a more detailed explanation on the evaluation campaign).

2.1 Evaluation description

The evaluation consists of segmenting a broadcast news audio document according to a series of audio classes. The proposed classes are speech, music, and noise. The classes can be present simultaneously in the audio signal (e.g. speech overlapped with music and/or background noise). Therefore, the aim is to indicate if one, two or the three classes are present for a given time instant. For this evaluation, speech is considered to be present if a person is speaking but not in the background, music is considered in a general sense, and noise is considered when any other sound -different than speech and music- is present.

Unlike the 2010 audio segmentation evaluation, there are no prior classes representing overlapped regions and a multiple layer labeling is proposed.

2.2 Database description

As in the last audio segmentation evaluation, the Catalan broadcast news database from the 3/24 TV channel is used to train systems in the present evaluation. The database was recorded by the TALP Research Center from the UPC, and it was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, includes around 87 hours of sound. The distribution of classes within the database is the following: speech is present a 92% of the time, music is present a 20% and noise in the background a 40%. Additionally, there is an *other* class which represents a 3% of the time. With regard to the overlapped regions, speech is found along with noise a 40% of the time, whereas speech over music is found a 15% of the time.

For the current evaluation, the Corporación Aragonesa de Radio y Televisión has donated part of the Aragón Radio archive and allows the use of these data for research purposes. Four hours of audio are used for development and 16 hours are used for testing.

All the audio signals are provided in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency.

2.3 Scoring

The proposed evaluation scoring is inspired in the method used in the NIST RT Diarization evaluations. Thus, the segmentation error score (SER) is computed as the fraction of class time that is not correctly assigned to that specific class (speech, noise or music). The SER includes the different kind of errors that can be found in a given segmentation:

- **Class Error Time:** Consists of the amount of time that has been assigned to an incorrect class
- **Missed Class Time:** Consists of the amount of time that a class is present but not labeled by the segmentation system.
- **False Alarm Class Time:** Consists of the amount of time that has been assigned to a class that is not really present.

In order to take into account both inconsistent human annotations and the uncertainty about class beginnings and ends, a forgiveness collar of one second before and after each reference boundary is considered.

3 System description

The system is based on GMMs. Figure 1 shows the overall architecture of the system. Each acoustic class is modeled by a 128 component, diagonal covariance GMM. For model training, only audio regions with no overlapping are considered (according to the provided labels). This way, one model is obtained for each class (speech, music, background noise).

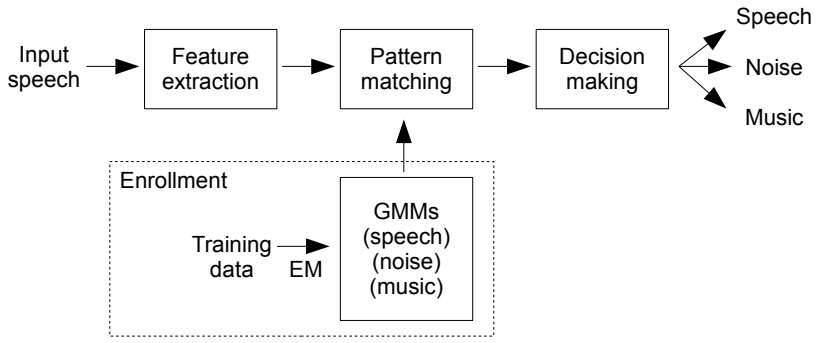


Fig. 1. Segmentation system architecture.

With regard to the used features, the standard MFCC method has been chosen, using the well-known 39-dimensional vectors, including delta and acceleration coefficients. One vector is extracted every 0.010 seconds with an overlap of 0.025 seconds. For feature extraction, HTK [3] software has been used. The

system has been developed using the Alize [1] library. It consists of an API which provides low-level functionality for develop *ad hoc* training and decoding programs. The training has been carried out through the expectation-maximization algorithm. using the 3/24 TV database, whereas the decoding process is based on log-likelihoods computation. For this particular task, a window of 1 second (i.e. 100 feature vectors) is used to calculate the mean log-likelihood. Then, the one second frame is assigned to the class whose model provides the highest log-likelihood. Next, log-likelihood subtractions between the remaining models are calculated and the presence of overlapping is decided according to a set of threshold values empirically determined. These threshold values were calculated in order to minimize the segmentation error rate. Therefore, a segment can be assigned to two or three classes according to their similarity for the given frame. The thresholds tuning was done using the first subset of the Aragon Radio database, provided as development data.

4 Results and discussion

The audio segmentation system described above has been tested on the second partition of the development set. Table 1 shows the results of the test. The overall segmentation error is shown along with the missed class, false alarm and class error times.

Table 1. Performance of the audio segmentation system: missed class time, false alarm class time, class error time and overall segmentation error, in seconds and percentages.

Error	time (s)	(%)
Missed class time	2979.82	25.00
False alarm class time	938.73	7.90
Class error time	488.70	4.10
Overall segmentation error	4407.25	**37.04**

Experiments have been performed on a Dell laptop with an Intel Core i5 2450M CPU at 2.5 GHz, and 4 GB of RAM. CPU times are shown in table 2, according to the feature extraction and audio segmentation tasks separately. Although the CPU is capable of 4 simultaneous threads, note that no parallelism is used for the experiments (i.e. the tasks are executed sequentially in only one CPU core).

Table 2. CPU time for feature extraction and audio segmentation tasks

Task	xRT
Feature extraction	0.0037
Audio segmentation	0.0040

Although the system presents good error rates for false alarms and class errors, the missed class time increases significantly up to a 25 %. A further analysis reveals the difficulty in the detection of background noise overlapped with speech.

5 Conclusions

An audio segmentation based on GMM, MFCC and decision making through log-likelihood thresholds has been developed and submitted to the Albayzin 2012 audio segmentation evaluation. On the one hand, the obtained results suggest that the GMM framework should be complemented with other methods in order to become more accurate for audio segmentation tasks. Particularly, determining the presence of background noise over speech is challenging. On the other hand, the technique is computationally efficient and may be useful for time critical applications. Additionally, the system performs adequately for speech activity detection.

References

1. J.F., B., F., W., S, M.: Alize, a free toolkit for speaker recognition. In: Proceedings of ICASSP 2005. Philadelphia, USA (2005)
2. Ortega, A., Castan, D., Miguel, A., Lleida, E.: The Albayzin 2012 Audio Segmentation Evaluation (2012), `http://iberspeech2012.ii.uam.es/images/PDFs/albayzin2012audiosegmentationevalplan.pdf`
3. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.4. Cambridge University Press (2009)

GTTS Systems for the Albayzin 2012 Audio Segmentation Evaluation*

Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, and Germán Bordel

GTTS (http://gtts.ehu.es), Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU
Barrio Sarriena, 48940 Leioa, Spain
luisjavier.rodriguez@ehu.es

Abstract. This paper briefly describes the audio segmentation systems presented by GTTS to the Albayzin 2012 Audio Segmentation Evaluation (ASE). The same systems presented to the Albayzin 2010 ASE have been applied in a blind fashion, that is, with no specific tunings to the Aragon radio archive recordings from which test signals have been extracted. The primary system consists of an ergodic Continuous Hidden Markov Model with 5 states and 512 mixtures per state, each state representing a mix of audio sources: (1) music, (2) clean speech, (3) speech with music in the background, (4) speech with noise in the background and (5) other (noise, long silence fragments, etc.). The emission distributions corresponding to the HMM states were estimated on segments extracted from the Catalan broadcast news (3/24 TV) database provided for development, and transition probabilities were heuristically fixed. Given an input signal, this model produces an optimal decoding (and segmentation) according to the maximum likelihood criterion. The contrastive system consists of five 1024-mixture Gaussian Mixture Models (one per class, for the five classes mentioned above), estimated independently using 3/24 TV segments and applied on a frame-by-frame basis to get a sequence of smoothed log-likelihoods. The class yielding the maximum likelihood is chosen at each frame, and finally a mode filter is applied to smooth the sequence of decisions. The output of both systems was filtered so that it consisted of 3-class (speech, music and noise) possibly overlapping segments, as required in this evaluation, assuming that at least one of the categories must appear at any given frame.

Index Terms: Audio Segmentation, Gaussian Mixture Models, Hidden Markov Models

* This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE11UN065), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds). Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

1 Introduction

Audio segmentation and classification is a preprocessing step required by many applications, typically to filter out signals coming from undesired sources. Our approach to audio segmentation was motivated by the need for a speech detector in a speaker diarization system, which would allow us to discard non-speech segments (containing music, noise, etc.), so that clustering was only performed on speech segments.

The two systems presented to this evaluation were originally developed for the Albayzin 2010 Audio Segmentation Evaluation (ASE), and therefore they deal with five types of audio sources: (1) music, (2) clean speech, (3) speech with music in the background, (4) speech with noise in the background and (5) other (noise, long silence fragments, etc.). The acoustic models (one per class) and the classification approaches have been built and tuned based exclusively on the 3/24 TV channel recordings distributed for development, without any adaptation/tuning to the Aragon radio archive recordings used as test signals in this evaluation. Note also that, according to the above described motivation, our systems were not optimized for audio classification but for speech detection.

The first approach applied a 5-class ergodic HMM and performed maximum-likelihood Viterbi decoding. This approach yielded quite good performance in the speech detection task. The second approach, based on GMM frame-by-frame scoring, was developed with the aim to improve performance on multiclass audio segmentation tasks. It yielded better results than the ergodic HMM system on the Albayzin 2010 ASE datasets, but it heavily relied on parameter tunings, so it is presented as contrastive system in this evaluation. In both cases, the system output was filtered in order to produce 3-class (speech, music and noise) possibly overlapping segments (in RTTM formatted files), as required in this evaluation. We assumed that at least one of the classes must appear at any given frame, so no frame was left unassigned. The mapping of classes is presented in Table 1.

Table 1. Mapping classes from Albayzin 2010 ASE into Albayzin 2012 ASE.

	Speech	**Music**	**Noise**
Music		√	
Clean speech	√		
Speech + Background Music	√	√	
Speech + Background Noise	√		√
Other			√

All speech processing, HMM/GMM estimation, Viterbi decoding and GMM likelihood computations were performed with the Sautrela toolkit [1]. Text processing and file manipulation were all performed by means of UNIX utilities and applications (awk, SoX, etc.).

2 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features. The choice of MFCC is based on the fact that historically there have been no features specifically designed for audio segmentation, and the MFCC are the most commonly used parameters for speech processing applications.

The audio was analysed in frames of 32 milliseconds (512 samples) at intervals of 10 milliseconds. A Hamming window was applied and a 512-point FFT computed. The FFT amplitudes were then averaged in 24 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. A Discrete Cosine Transform was finally applied to the logarithm of the filter amplitudes, obtaining 13 Mel-Frequency Cepstral Coefficients (MFCC), including the zero (energy) coefficient. Cepstral Mean Subtraction (CMS) was not applied, in order to keep channel and background information that may be relevant for audio classification.

3 Audio segmentation based on HMM decoding (primary system)

For development purposes, only the first 16 sessions of the 3/24 TV channel dataset were used. Sessions 3, 7, 11 and 13 were used for tuning purposes. The remaining 12 sessions were used to estimate model parameters, by splitting them (by means of SoX) into five subsets of segments (one per class), according to reference segmentations. A single-state HMM was estimated for each class, using the Baum-Welch algorithm on the corresponding set of segments. An ergodic Continuous Hidden Markov Model was then built by composing the five single-state HMMs under the Layered Markov Model framework defined in Sautrela [2]. Given an input sequence of feature vectors, the optimal decoding (and segmentation) was obtained by applying the Viterbi algorithm to get the optimal sequence of states in the ergodic HMM.

The number of mixtures per state (512) and the transition probabilities (auto-transitions fixed to 0.999999, transitions between states and final state transitions fixed to $2 \cdot 10^{-7}$) were optimized on audio segmentation experiments over the 4 tuning sessions mentioned above. Though system performance was quite poor for the multiclass setup defined in the Albayzin 2010 ASE, when applying it under a speech detection setup, the false alarm error rate was 1.16% and the miss error rate was 1.78% for the speech class.

4 Audio segmentation based on frame-by-frame GMM scoring (contrastive system)

This system was also developed using 16 sessions of the 3/24 TV channel dataset, as for the HMM-based system (12 sessions for training, 4 sessions for tuning). A GMM was estimated for each class, starting from the corresponding subset of

training segments. Given an input sequence of feature vectors, the set of GMMs was applied to compute frame-by-frame log-likelihoods. A smoothing window of length N was then applied, so that the log-likelihood of each class i was replaced by the arithmetic mean computed in that window, as follows:

$$\hat{ll}(i,t) = \frac{1}{N} \sum_{k=-N/2}^{N/2} ll(i,t+k)$$

At each frame, the class yielding the highest smoothed likelihood was chosen, and a frame-level sequence of class labels was produced. Finally, a mode filter of length M was applied to smooth the sequence of decisions. The number of mixtures of the GMMs (1024), the length of the score smoothing window ($N = 100$) and the length of the mode filter ($M = 200$) were optimized on audio segmentation experiments over the 4 tuning sessions. This system yielded better results than the HMM-based system in both the multiclass setup defined in the Albayzin 2010 ASE and the speech detection setup on which we were interested (false alarm error rate = 1.14% and miss error rate = 1.32%, for the speech class). However, since the performance of this system was quite sensitive to parameter tunings (and these were not optimized for the Aragon radio archive recordings from which test signals have been extracted in the Albayzin 2012 ASE), we have presented it as contrastive system.

5 Results on the Aragon radio development datasets

Tables 2 and 3 show the performance of the two audio segmentation systems described above on the Aragon radio development datasets (dev1 and dev2) provided for this evaluation. Besides the segmentation error score in the 3-class audio segmentation task, the segmentation error score in the corresponding speech detection task (computed by considering only the speech segments in both reference and system segmentations) is shown too. In both cases, miss, false alarm and class labeling error rates are presented in parentheses. Finally, to provide a performance ground, we provide the segmentation error scores for a *trivial* system which outputs a single segment including the full signal as containing both speech and music (the most common classes).

Table 2. Performance of the GTTS primary and contrastive audio segmentation systems on the Aragon radio development dataset *dev1*.

	% SER (miss, false alarm, labeling error)	
	3-class audio segmentation task	**speech detection task**
Primary	35.45 (12.8, 8.9, 13.8)	3.02 (2.5, 0.5, 0.0)
Contrastive	34.94 (11.8, 8.9, 14.2)	1.96 (1.1, 0.9, 0.0)
Trivial	35.74 (4.2, 15.6, 16.0)	7.33 (0.0, 7.3, 0.0)

Table 3. Performance of the GTTS primary and contrastive audio segmentation systems on the Aragon radio development dataset *dev2*.

	% SER (miss, false alarm, labeling error)	
	3-class audio segmentation task	**speech detection task**
Primary	38.70 (17.8, 9.4, 11.6)	3.03 (1.2, 1.8, 0.0)
Contrastive	37.29 (16.3, 8.8, 12.1)	2.83 (0.3, 2.5, 0.0)
Trivial	41.26 (3.4, 20.6, 17.2)	11.44 (0.0, 11.4, 0.0)

Experiments were carried out on a Dell PowerEdge 1950, equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 4GB of RAM. CPU times (in terms of real-time factor, $\times$RT) are shown in Table 4, considering three separate operations: (1) feature extraction, (2) model estimation and (3) audio segmentation. In the latter case, I/O operations and all the secondary computations needed to carry out the audio segmentation task are counted. Note that the primary system employs more time than the contrastive system for model estimation, but is faster for audio segmentation, providing only slightly worse performance in the speech detection task. The total CPU time, computed by adding CPU times for feature extraction and audio segmentation, falls below 0.05$\times$RT in both cases.

Table 4. CPU time (real-time factor, $\times$RT) employed in feature extraction, model estimation and audio segmentation for the GTTS primary and contrastive systems.

	Primary	**Contrastive**
Feature extraction	0.0033	
Model estimation	0.4819	0.1205
Audio segmentation	0.0375	0.0458

6 Conclusions

For the Albayzin 2012 Audio Segmentation Evaluation, GTTS has applied two fast systems (CPU time requirements falling below 0.05$\times$RT): a primary system based on a five-class ergodic HMM which outputs the optimal Viterbi-based sequence of states (and thus of classes) given an input signal, and a contrastive system based on frame-by-frame GMM scoring, followed by smoothing and mode filtering. The development effort has been almost null, since we simply recycled the two systems developed for the Albayzin 2010 Audio Segmentation Evaluation, applied them to the Aragon radio development and test signals (with no tunings), and adapted the output to the 3-class RTTM formatted output required in this evaluation, assuming that speech, music and/or noise must appear at any given frame.

When applied to the 3-class audio segmentation task, error rates were above 35%, only slightly better that those attained by a trivial system providing a fixed *speech and music* output. On the other hand, the same systems provided remarkably good performance in the corresponding speech detection task (between 2% and 3% detection error rates). This makes them suitable as speech detectors in many applications, such as speaker diarization, language recognition, ASR, etc.

References

1. M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.
2. M. Penagarikano and G. Bordel, "Layered Markov Models: A New Architectural Approach to Automatic Speech Recognition," in *Proceedings of the MLSP Workshop*, (Sao Luis, Brasil), pp. 305–314, October 2004.

Albayzin Evaluation: The PRHLT-UPV Audio Segmentation System

J. A. Silvestre-Cerdà, A. Giménez, J. Andrés-Ferrer, J. Civera, and A. Juan

Universitat Politècnica de València, Camí de Vera s/n,
46022 València, Spain,
jsilvestre@dsic.upv.es,
http://translectures.eu

Abstract. This paper describes the audio segmentation system developed by the PRHLT research group at the UPV for the Albayzin Audio Segmentation Evaluation 2012. The PRHLT-UPV audio segmentation system is based on a conventional GMM-HMM speech recognition approach in which the vocabulary set is defined by the power set of segment classes. MFCC features were extracted to represent the acoustic signal and the AK toolkit was used for both, training acoustic models and performing audio segmentation. Experimental results reveals that our system provides an excellent performance on speech detection, so it could be successfully employed to provide speech segments to a diarization or speech recognition system.

1 Introduction

Audio segmentation is a task with applications in subtitling, content indexing and analysis that has received notable attention due to the increasing application of automatic speech recognition (ASR) systems to multimedia repositories and broadcast news [6–9]. Formally, this task can be stated as the segmentation of a continuous audio stream into acoustically homogenous regions. Audio segmentation facilitates posterior speech processing steps such as speaker diarization or speech recognition.

Previous work on audio segmentation can be classified into those tackling this task at the feature extraction level [1, 10, 4, 3], and those approximations working at the classification level [2, 5]. This latter approximation is adopted in our audio segmentation system.

This paper describes the PRHLT-UPV audio segmentation system. First, the Albayzin Audio Segmentation Evaluation 2012 is presented in Section 2. Next, a complete system description is provided in Section 3. Experimental results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 Albayzin 2012 audio segmentation evaluation

2.1 Database description

The database used for the audio segmentation evaluation consists of a Catalan broadcast news database from the 3/24 TV channel, which comprises 87 hours of acoustic data for training purposes. In this dataset, speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called others was defined which can be found a 3% of the time. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music. Table 1 shows the audio time distribution over all overlapping acoustic classes as disjoint sets for the training set.

In addition, two sets, *dev1* and *dev2*, from the Aragón Radio database of the Corporación Aragonesa de Radio y Televisión (CARTV), are used for developing and internal testing purposes, respectively. Both sets sums up to 4 hours of acoustic data. All audio signals are provided in PCM format, mono, little endian 16 bit resolution, and sampling frequency of 16 kHz.

Table 1. Audio time distribution of all overlapping classes for the training set.

Class	Time (h)	Time (%)
sp	31.85	38.2
mu	4.94	5.9
no	0.91	1.1
sp+mu	12.58	15.1
sp+no	31.36	37.6
no+mu	0.06	0.1
sp+no+mu	1.65	2.0
Total	87	100

2.2 Evaluation metric

In order to assess the quality of our system, we used the Segmentation Error Rate metric (SER), defined as the fraction of class time that is not correctly attributed to that specific class (speech, noise or music):

$$SER = \frac{\sum_n T(n)\left[max(N_{ref}(n), N_{sys}(n),) - N_{correct}(n)\right]}{\sum_n T(n) N_{ref}(n)} \quad (1)$$

where $T(n)$ is the duration of segment n, $N_{ref}(n)$ is the number of reference classes that are present in segment n, $N_{sys}(n)$ is the number of system classes that are present in segment n, and $N_{correct}(n)$ is the number of reference classes in segment n correctly assigned by the segmentation system.

A forgiveness collar of one second, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a class begins or ends.

3 System description

Audio segmentation can be viewed as a simplified case of ASR, in which the system vocabulary is constituted by the power set of segment classes: Speech (sp), music (mu) and noise (no). For the Albayzin evaluation the silence (si) class is also included to denote that none of the three classes is present in a given time instant. Thus, the system vocabulary is defined as

$$\mathcal{C} = \{sp, mu, no, sp + mu, sp + no, mu + no, sp + mu + no, si\} \tag{2}$$

Provided an audio stream $\boldsymbol{x}$, the segmentation problem can be stated from a statistical point of view as the search of a sequence of class labels $\hat{\boldsymbol{c}}$ so that

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c} \in \mathcal{C}^*}{\operatorname{argmax}}\, p(\boldsymbol{x} \mid \boldsymbol{c})\, p(\boldsymbol{c}) \tag{3}$$

where, as in ASR, $p(\boldsymbol{x} \mid \boldsymbol{c})$ and $p(\boldsymbol{c})$ are modelled by acoustic and language models, respectively. In our case, it should be noted that each word is composed by a single phonem.

Acoustic models were trained on MFCC feature vectors computed from acoustic samples using the HTK HCopy tool. We used a 0.97 coefficient pre-emphasis filter and a 25 ms Hamming window that moves every 10 ms over the acoustic signal. From each 10ms frame, a feature vector of 12 MFCC coefficients is obtained using a 26 channel filter bank. Finally, the energy coefficient and the first and second time derivatives of the cepstrum coefficients are added to the feature vector.

Each segment class is represented by a single-state Hidden Markov Model (HMM) without loops, and its emission probability is modelled by a Gaussian Mixture Model (GMM). Acoustic HMM-GMM models were trained using the AK toolkit[1], which implements the conventional Baum-Welch algorithm. For each segment class, the number of mixture components per state was tuned on the development set. A 5-gram back-off language model with constant discount was trained on the sequence of class labels using the SRILM toolkit [11]. Constant discounts for each order were optimised on the development set. The segmentation process (search) was also carried out by the AK toolkit.

4 Experimental results

This section is devoted to the description of the experimental setup and results performed before submitting our final audio segmentation system. For these

[1] http://sourceforge.net/projects/aktoolkit

experiments, acoustic and language models were trained on the *training* set, while acoustic and language model parameters were tuned on the *dev1* set.

Table 2 shows SER figures computed on the *dev1* and *dev2* sets. In addition to the overall SER, SER values are provided for each acoustic class (speech, noise, music) in isolation. As observed in Table 2, our audio segmentation system offers an excellent performance in speech detection, so it could be successfully employed to provide speech segments to diarization or speech recognition systems.

Table 2. Segmentation error rate (SER) for the three acoustic classes (speech, music, noise) in isolation and overall SER, computed over the *dev1* and *dev2* sets.

Set	Speech	Music	Noise	Overall
dev1	1.2	25.3	71.4	24.9
dev2	2.2	20.2	71.2	26.4

However, the system provides low performance at detecting non-speech classes, specially the noise class. This fact can be explained by two reasons. First, we are using a feature representation of the acoustic signal that is focused on highlighting human voice characteristics, and conversely to abate acoustic features from music and noise. Secondly, few music and noise data samples appear in isolation (5% and 1%, respectively) to make feasible to robustly estimate acoustic models for these classes. For this reason, the global classifier suffers from a bias towards the isolated speech class. For instance, the posterior probability of an *sp+no* segment, given the isolated speech model parameter set is expected to be larger than that of the isolated noise model parameter set.

5 Conclusions

This paper has described the PRHLT-UPV audio segmentation system for the Albayzin evaluation. This system tackles the task of audio segmentation from the viewpoint of ASR system with a reduced vocabulary set. The vocabulary set comprises the speech, music and noise classes and combinations of those classes defined for this task. The experimental results show that this system provides excellent performance detecting speech segments, but its performance decays when dealing with music or noise segments.

References

1. Ajmera, J., McCowan, I., Bourlard, H.: Speech/music segmentation using entropy and dynamism features in a hmm classification framework. Speech Communication 40(3), 351–363 (2003)
2. Bugatti, A., Flammini, A., Migliorati, P.: Audio classification in speech and music: A comparison between a statistical and a neural approach. EURASIP J. Audio Speech Music Process. pp. 372–378 (2002)

3. Gallardo-Antolín, A., Montero, J.M.: Histogram equalization-based features for speech, music, and song discrimination. IEEE Signal processing letters 17(7), 659–662 (2010)
4. Izumitani, T., Mukai, R., Kashino, K.: A background music detection method based on robust feature extraction. In: Proc. of ICASSP. pp. 13–16
5. Lavner, Y., Ruinskiy, D.: A decision-tree-based algorithm for speech/music classification and segmentation. EURASIP J. Audio Speech Music Process. 2009, 2:1–2:14 (2009)
6. Li, D., Sethi, I.K., Dimitrova, N., McGee, T.: Classification of general audio data for content-based retrieval. Pattern Recognition Letters 22(5), 533 – 544 (2001)
7. Lu, L., Jiang, H., Zhang, H.: A robust audio classification and segmentation method. In: Proc. of ACM International Conference on Multimedia. pp. 203–211 (2001)
8. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: Proc. of ICASSP. vol. 2, pp. 5–8 (2003)
9. Nwe, T., Li, H.: Broadcast news segmentation by audio type analysis. In: Proc. of ICASSP. vol. 2, pp. 1065–1068 (2005)
10. Panagiotakis, C., Tziritas, G.: A speech/music discriminator based on rms and zero-crossings. IEEE Transactions on Multimedia 7(1), 155–166 (2005)
11. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP'02. pp. 901–904 (September 2002)

The UVigo-GTM System for the Albayzin'12 Audio Segmentation Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
Universidade de Vigo
E.E. Telecomunicación, 36310 Vigo (Spain)
{plopez,ldocio,carmen}@gts.uvigo.es

Abstract. This paper describes the UVigo-GTM system for the Albayzin 2012 Audio Segmentation Evaluation. The system can be divided in two stages, segmentation and classification. Segmentation is performed by a Bayesian Information Criterion (BIC)-based algorithm where a constraint is added: a change-point has to fit a stochastic Poisson process, in other case it will be rejected. The classification stage involves two different classifiers, one based on a log-likelihood detector and the other one based on support vector machines, whose outputs are fused. The segmentation strategy showed better results than the baseline BIC strategy, while the fusion of the two classifiers performed better than any of them individually.

Keywords: Audio segmentation, Bayesian information criterion, log-likelihood classification, support vector machines, fusion of classifiers.

1 Introduction

Audio segmentation is a task consisting in dividing an audio stream into homogeneous regions according to some criteria. In the Albayzin 2012 Audio Segmentation Evaluation, an audio signal has to be divided into segments that include speech, music, noise or any combination of these three classes. This paper describes an audio segmentation system that was implemented for this evaluation. It has two stages: the segmentation stage, which divides the audio signal into homogeneous regions, and the classification stage, that assigns a class to each segment.

Bayesian Information Criterion (BIC) is a widely used approach for audio segmentation [3]. It is fast and has an acceptable performance, but one of its problems is the high number of insertions (the algorithm detects change-points that are not actual change-points). The segmentation algorithm described in this paper introduces an approach to reduce the number of insertions caused by the BIC algorithm. This is done by adding a new constraint to the segmentation algorithm, which leans on the idea that the occurrence of change-points on an audio stream can be modeled by means of a stochastic Poisson process [8]. Thus, anytime BIC finds a change-point, this point is only accepted if it fits a Poisson process.

Classification is performed using an approach that fuses two different classification systems. One of these systems computes the log-likelihood of a segment given the models of the different classes. The other system models the audio segment by means of a supervector and classify it with a support vector machine, obtaining the probability estimates for each of the possible classes. Next, a backend is applied to the scores that are obtained by each classification system, and these new scores are fused in order to obtain a log-likelihood for each class that is a combination of the scores of the two classification systems. The class whose model achieves the highest log-likelihood is the one that is selected for the audio segment.

The paper is organized as follows: section 2 describes the audio segmentation system in general terms; section 3 describes the features that are used in the system; the segmentation approach is described in section 4; section 5 explains the two classification systems, its fusion and the decision-making; a description of the experiments is given in section 6; and section 7 summarizes the experimental results.

2 System Overview

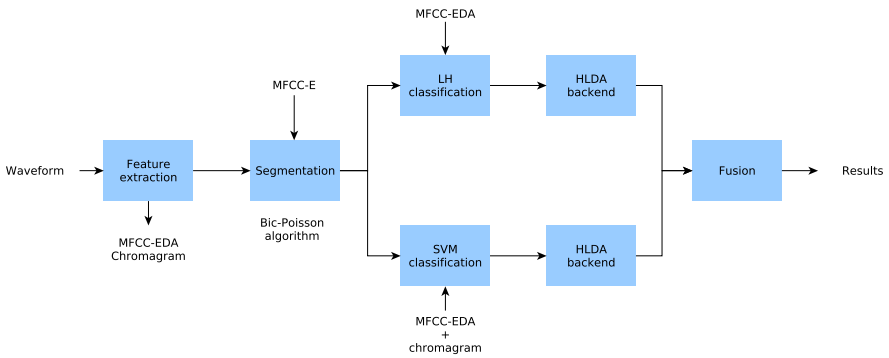


Fig. 1. System overview

Figure 1 shows a block diagram of the audio segmentation system. First, features are extracted from the waveform, and once these features are obtained the audio file is segmented, obtaining a set of segments. Each segment has to be classified into one of the seven classes that are defined: speech (s00), music (0m0), noise (00n), speech with music (sm0), speech with noise (s0n), music with noise (0mn) and speech with music and noise (smn). Two different classifiers are used, one based on log-likelihoods (LH) and the other one based on support vector

machines (SVM), obtaining two sets of seven scores, one per class. A backend is applied to each of these set of scores, and then they are fused, obtaining one log-likelihood per class: the class assigned to the segment is the one that obtains the highest log-likelihood. In the following sections, each block is described in more detail.

3 Feature extraction

Two types of features are used in this audio segmentation system. 12 Mel-Frequency Cepstral Coefficients (MFCC) are extracted using a 25ms Hamming window at a rate of 10ms per frame, and augmented with the normalized log-energy and their delta and acceleration coefficients. Also, the chromagram of the waveform is extracted; a chromagram is a sequence of 12-dimensional vectors that describe the relative energy of the twelve pitch classes of the western tonal scale [1]. The chromagram is also extracted at a rate of 10ms per frame.

In the segmentation stage only the 12 MFCCs and the log-energy are used; in LH classification the 39 MFCCs are used (MFCC-EDA); and in SVM classification the 39 MFCCs and the chromagram are concatenated, obtaining vectors of 51 features.

4 Segmentation

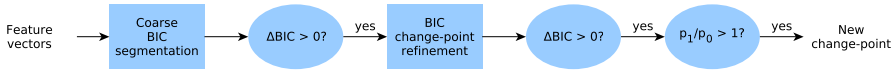


Fig. 2. Segmentation strategy

The BIC criterion is a hypothesis test to decide whether there is a change-point in a window of data (H_1) or not (H_0) by observing a value ΔBIC: if ΔBIC is greater than zero then H_1 is true (there is a change-point in the window), otherwise there is no change-point and H_0 is true [3].

A Bayesian Information Criterion (BIC) system as described in [3] is implemented to perform speaker segmentation: a window of data that slides and grows is analyzed in order to detect a possible speaker change in it by applying the BIC criterion [7]; anytime a change-point is found, a fixed-size window is centered on this change-point and the BIC criterion is applied again in order to refine its position or to discard it. If the change-point is discarded, the window will grow or slide, depending on the size reached by the window, and the procedure is repeated again.

The BIC segmentation algorithm has a parameter λ to tune the penalty applied to the hypothesis test [3]. Depending on the value of λ, the system will have a higher number of false alarms or a higher number of miss detections, so it has to be tuned in order to find a trade-off between these two types of errors.

A technique to reduce the number of false alarms was implemented on the BIC algorithm. In this strategy, it is assumed that the change-point instants follow a Poisson process. A Poisson process is an independent occurrence process where the number of occurrences in two disjoint time intervals is independent, the probability of an occurrence is proportional to the observed interval and occurrences are not simultaneous [2].

An homogeneous Poisson process is characterized by its rate $\lambda = \frac{1}{\tau}$, where τ is the mean time between occurrences. This parameter is easily estimated by computing the mean of the segment durations on the training data. In this type of processes, the probability of not having an occurrence in a time interval t and the probability of having one or more occurrences in a time interval t can be computed as follows:

$$p_0(t) = e^{-\lambda t} \tag{1}$$

$$p_n(t) = 1 - p_0 = 1 - e^{-\lambda t} \tag{2}$$

Anytime a change-point is detected by the BIC algorithm, which forms a segment of duration t, $p_0(t)$ and $p_n(t)$ are computed and the following constraint is applied:

$$\frac{p_n(t)}{p_0(t)} \begin{cases} > 1 \rightarrow \text{accept change-point} \\ < 1 \rightarrow \text{reject change-point} \end{cases} \tag{3}$$

This means that, given a Poisson process with rate λ that models the arrival of change-points, if the probability of not having a change-point in a time interval t is less than the probability of having one or more change-points in a time interval t, the change-point is considered as an insertion and, therefore, rejected. In other case, the change-point is accepted.

Figure 2 shows a block diagram of the proposed segmentation system.

5 Classification

The two classification strategies and their fusion are depicted in figure 3. In this section, this whole procedure is described in more detail.

5.1 LH Classification

Gaussian mixture models (GMMs) Θ_i of 32 mixtures are trained for each of the seven classes i defined in Sec. 2. The reason why the number of mixtures is 32 Gaussians is because there is little training data. After the audio stream is segmented, for each speech segment S, the probability of having S given each of the seven models is computed. Thus, a set of log-likelihoods $\{LH(S|\Theta_i)\}, i \in (s00, 0m0, 00n, sm0, s0n, 0mn, smn)$ is obtained [6].

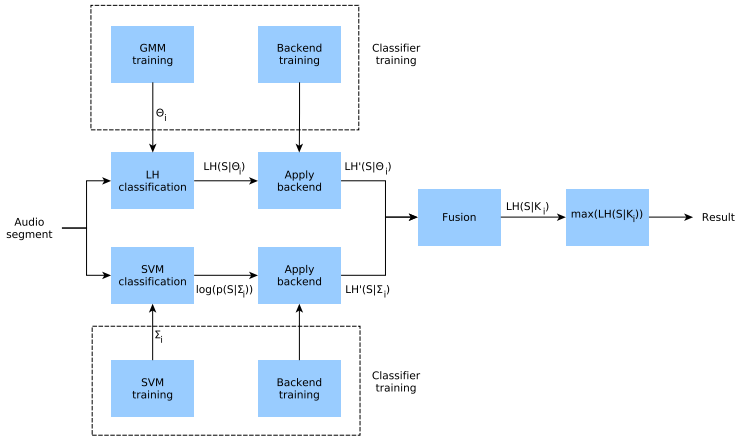


Fig. 3. Classification strategy

5.2 SVM Classification

A universal background model (UBM) is adapted to manually labeled training data obtaining a set of GMMs, whose means are concatenated in order to obtain a set of supervectors. These supervectors are used to train a support vector machine that discriminates among the seven classes defined above. Once the SVM is trained, the segments obtained from the segmentation stage are classified. For this, the UBM is adapted to each audio segment. Then, the supervectors obtained by concatenating the means of the resulting GMMs are classified by the SVM. Thus, for each segment, a set of seven log-probabilities $\{LH(S|\Sigma_i)\}, i \in (s00, 0m0, 00n, sm0, s0n, 0mn, smn)$ is obtained. SVM training and classification are performed using the library for large linear classification LIBLINEAR [4].

5.3 Backends and Fusion

The classification algorithms described in Sect. 5.1 and 5.2 output two sets of seven scores (one per class) for each audio segment. A heteroscedastic linear discriminant analysis (HLDA) backend is applied to each system in order to calibrate these scores (log-likelihoods in the LH classification and log-probabilities in the SVM classification). Once the two sets of calibrated scores are obtained a discriminative fusion is applied, obtaining a new set of seven scores (log-likelihoods). The class that achieves the highest log-likelihood is selected as the class that has to be assigned to the audio segment. Backends and fusion were applied using the FoCal multiclass toolkit [5].

6 Description of the Experiments

6.1 Evaluation Metrics

To evaluate the performance of the audio segmentation system, the segmentation error time (SER) is considered. This metric is a combination of the class error time (CET), the missed class time (MCT) and the false alarm class time (FACT).

6.2 Experiments

The data provided for the evaluation is divided into two datasets, 324TV and AragonRadio. Aragon Radio is also divided in dev1, dev2 and test. A UBM was trained with the 324TV dataset; this UBM was used in the SVM classification stage to obtain the supervectors for the segments.

Two different experiments were run in order to avoid bias, and the datasets used in each of them are described in table 1. The train dataset of each experiment is used for training the GMMs used for LH classification, for training the SVM in the SVM classification stage, and for training the backends and the fusion parameters.

Table 1. Datasets of the experiments

Experiment	Train	Test
1	dev1	dev2
2	dev2	dev1

Once the two experiments are run, the results obtained in the test stage of the two experiments are joined together in order to obtain the evaluation metrics of the whole development dataset. The free parameters of the system (λ in the segmentation stage and the number of Gaussians M of the UBM in the SVM classification stage) are chosen by selecting the ones that achieve the lowest SER on development data. In these experiments, the selected values were $\lambda = 3.5$ and $M = 64$.

7 Experimental Results

Table 7 shows the results obtained on the experiments described in Sect. 6 with the tuned parameters. The column labeled as "All" shows the results obtained when joining the outputs of experiments 1 and 2; columns 1 and 2 show the results obtained on the two test datasets defined in table 1.

Table 2. Experimental results on development set

Experiment	MCT	FACT	CET	SER
1	5.9%	11.6%	2.8%	20.27%
2	8.1%	3.5%	8.7%	20.20%
All	7.2%	6.9%	6.2%	20.23%

Acknowledgements This work has been supported by the Galician Regional Government (CN2011/019, 2009/062), Spanish Government (TEC009-14094-C04-04 and FPI grant BES-2010-033358), and the European Regional Development Fund.

References

1. Ahonen, T. E., Compression-Based Clustering of Chromagram Data: New Method and Representations, In: Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval, pp. 474–481 (2012).
2. Allen, A. O., Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 2nd edition (1990).
3. Cettolo, M. and Vescovi, M., Efficient Audio Segmentation Algorithms Based on the BIC, In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. VI, pp. 537–540 (2003).
4. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J., LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research, vol. 9, 1871–1874 (2008). Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`
5. FoCal, toolkit for evaluation, fusion and calibration of statistical pattern recognizers (2008), `http://sites.google.com/site/nikobrummer/focal`
6. Reynolds, D., Quatier, T., Dunn, R., Speaker Verification Using Adapted Gaussian Mixture Models Digital Signal Processing, vol. 10, 19–41 (2000).
7. Schwarz, G., Estimating the Dimension of a Model, Annals of Statistics vol. 6, n. 2, pp. 461–464 (1978).
8. Vinciarelli, A., Sociometry Based Multiparty Audio Recordings Segmentation, In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2006).

Albayzin Search-on-Speech Evaluation

System Descriptions

IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop

The ELiRF Query-by-Example STD systems for the Albayzin 2012 Search on Speech Evaluation

Emilio Sanchis, Lluís-F. Hurtado, Jon A. Gómez,
Marcos Calvo, and Raül Fabra

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{esanchis,lhurtado,jon,mcalvo}@dsic.upv.es
rafabbo@fiv.upv.es

Abstract. In this paper, we present two approaches to Query-by-Example Spoken Term Detection. In both systems a first phase that obtains posterior probabilities of phonemes is performed. One of the systems performs a Segmental DTW algorithm to obtain the best matchings between the query utterance and the speech data. The other system is based on a search algorithm that finds the best matchings between two graphs of phonemes that represent the query utterance and the speech data.

Keywords: Spoken Term Detection, Query-by-Example, Automatic Speech Recognition

1 Introduction

In recent years some approaches for mining in audio repositories have been developed. Classification, language identification, diarization, indexing or spoken term detection are some of these works. In particular, the Query-by-Example Spoken Term Detection is an interesting task, where it is not necessary to have any a-priori knowledge about the language corresponding to de audio. The search is based only in utterances of the words to search. Due to the problem of lacking the vocabulary of the task, many approaches are based on a phonetic representation of the speech or work directly on the parametrization of the signal [1–4].

In the work presented in this paper, we propose two approaches for the Query-by-Example Spoken Term Detection task. Both are based on a first phase where the posterior phonetic probabilities for each frame are obtained. This phonetic probabilities are computed from a previous process of acoustic clustering and classification in terms of acoustic classes. Once we have obtained the phonetic probabilities, in one of the systems we apply a Segmental Dynamic Time Warping algorithm to find the segments of the signal that best match the utterance. In the other system, we construct graphs of phonemes, with their corresponding probabilities, to represent the query utterance and the speech data. Then we have developed an algorithm to obtain the segments of speech that best match the graph representing the query utterance.

2 Computation of the *a posteriori* probabilities for the phonemes

The *a posteriori* probabilities of phonemes are calculated by combining probabilities of acoustic classes, which are obtained from a clustering procedure on the acoustic feature space, and the conditional probabilities of each acoustic class with respect to each phonetic unit [5].

In the clustering procedure, it is assumed that acoustic classes can be modelled by means of Gaussian distributions. Parameters of each Gaussian distribution are estimated by using the unsupervised version of the Maximum Likelihood Estimation (MLE) procedure. Thus, it is possible to estimate $\Pr(a|x_t)$, that is, the probability of each acoustic class a from the set A of acoustic classes, given an acoustic frame x at time t, x_t, from the GMM. Nevertheless, as we need the probability of each phonetic unit u from the set U of phonetic units, given an acoustic vector x_t, $\Pr(u|x_t)$, a set of conditional probabilities are estimated in order to calculate the phonetic probabilities from the acoustic ones.

The use of conditional probabilities allows us to compute the phonetic-conditional probability density $p(x_t|u)$ as follows [6]:

$$p(x_t|u) = \sum_{a \in A} p(x_t|a) \cdot \Pr(a|u) \tag{1}$$

for each $u \in U$, where $p(x_t|a)$ is the acoustic class-conditional probability density, computed as a Gaussian probability density function, and $\Pr(a|u)$ is the conditional probability that acoustic class a has been manifested when phonetic unit u has been uttered. Then, applying the Bayes formulation, we obtain the phonetic probabilities as:

$$\Pr(u|x_t) = \frac{\sum\limits_{a \in A} p(x_t|a) \cdot \Pr(a|u)}{\sum\limits_{v \in U} \Big(\sum\limits_{a \in A} p(x_t|a) \cdot \Pr(a|v) \Big)} \tag{2}$$

for each $u \in U$.

The set of conditional probabilities $\Pr(a|u)$ for all $a \in A$ is initially computed from a coarse segmentation [6]: An iterative process updates the conditional probabilities until no improvements on segmentation are found.

In order to estimate these acoustic probabilities we have used the training subcorpus of the Albayzin database [7].

3 The ELiRF-PhGraph QbE-STD system

To perform the Query-by-Example Spoken Term Detection task, one of the systems we have developed is based on graphs of phonemes. In this approach, first we generate a graph of phonemes per query and document in the collection. These graphs of phonemes are a compact representation of the information about the

phonemes that may have been uttered between some specific instants. Then, a search algorithm is run in order to compute the candidate detections of a query in a given audio document. Finally, these candidate detections are filtered according to their score, obtaining in this way the final detections.

3.1 Generation of the graphs of phonemes

The constructor of graphs of phonemes can be seen as a module for the ASR system that takes as input vectors of phonetic probabilities, generated by the Acoustic-Phonetic Decoder, and computes phonemes graphs as output. These graphs have a left-to-right topology, as shown in Figure 1, so they are directed and acyclic.

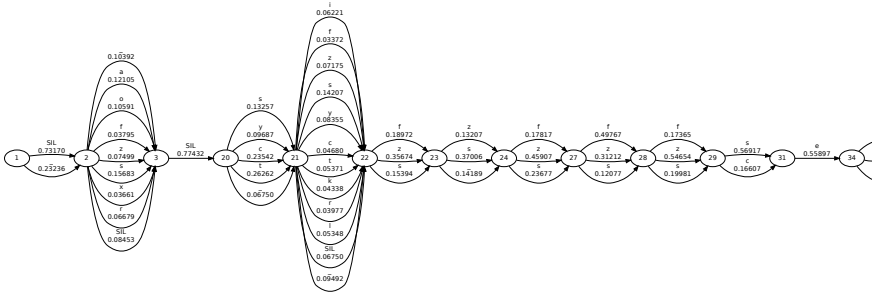


Fig. 1. Initial 340 ms extracted from a real example of phonemes graph.

Nodes act like time marks, as every node has a timestamp. A new node is created whenever the detection of any phonetic unit begins or ends, but normally the lapse of time in which a phonetic unit is detected comprises several arcs, i.e., given a detected phonetic unit that starts at one node and ends at a later one, other intermediate nodes could have been inserted between them. Regarding arcs, they have associated the phonetic unit uttered between the timestamps kept by origin and destination nodes, and also its phonetic probability.

The construction algorithm works with two thresholds: one for detection and another for extension. The first one is used to detect if any phoneme has been pronounced on a given frame, whereas the second one is used to determine when the pronunciation of any phoneme starts or finishes.

This algorithm has three steps: phoneme detection, error correction and hypotheses activation. In the first step, for each vector of phonetic probabilities, is analyzed in order to find if there is any probability above the detection threshold. In this case, we consider that a phoneme has been uttered, but the time in which the pronunciation started is still unknown, as well as the time it finishes. The extension threshold is used for solving this problem. When a pronunciation

is detected, then it starts a backward extension, which consists on marking as detected previous frames while they overpass the extension threshold. This way, the starting frame of the phoneme is determined. A similar process is used to find the ending node.

Next, in order to avoid undesirable effects due to unfiltered noise or other events, an error correcting phase is performed. In this phase spurious detections of phonemes are detected, as well as spurious misses of a phoneme in a zone where it is well detected.

Finally, the graph of phonemes is built according to these corrected detections. In this graph, there is a node in the graph representing an instant if there is a detected unit that starts or ends at that time. Also there are only arcs from one node to the following one. Each of these arcs represent a detection of a phoneme or a part of it, and have attached to them the detected phoneme. The weight of each arc is the accumulated probability of the detection between the instants represented by the starting and ending nodes.

3.2 Search algorithm

Our search algorithm is based on the idea that if a path on the graph corresponding to the query is contained in the graph of a document between the nodes s and e, then it is likely that the word expressed by the query is contained in the document, between the instants corresponding to the timestamps of s and e. The score of this occurence is the sum of the weights of both paths individually.[1] If, given two nodes in the document s and e, several paths that match other paths in the full query are found, only the occurence with maximum weight is considered. Also, edit operations (substitution, insertion and deletion of arcs) are allowed, in order to make this search more flexible and correct possible errors done while building the graph of phonemes. Insertion and deletion are penalized by means of a constant, while the cost of a substitution depends on the pair of phonemes being considered. Thus, matching two arcs is seen as a substitution where the phonemes attached to both arcs are the same.

The algorithm that searches for these occurences follows a Dynamic Programming strategy. Let M be a matrix of dimensions $I \times J$, where I and J are the number of nodes of the graphs representing the document and the query, respectively. So, $M(i,j)$ will contain the best score of arriving to node i in the document and j in the query by means of the edit operations. Also, given an arc a, let ori(a) and dest(a) be functions that return respectively the position in the graph of the starting and ending nodes of a, W(a) a function that returns the weight of the arc and S(a) a function that provides the symbol (phoneme) attached to the arc. Thus, formally, the algorithm can be expressed as follows:

$$M(i,j) = \begin{cases} 0 & \text{if } j = 0 \\ \max\{arcSub(i,j), arcIns(i,j), arcDel(i,j)\} & \text{otherwise} \end{cases} \quad (3)$$

[1] The sum is used because the weights represent logarithms of the *a posteriori* probabilities of the phonemes.

where:

$$arcSub(i,j) = \max_{\substack{\text{arcs } a,b: \\ \text{dest}(a)=i \ \wedge \ \text{dest}(b)=j}} \{M(\text{ori}(a),\text{ori}(b)) + \text{W}(a) + \text{W}(b) + k_{sub}(\text{S}(a),\text{S}(b))\}$$

$$arcIns(i,j) = \max_{\forall \text{ arc } a\text{: } \text{dest}(a)=i} \{M(\text{ori}(a),j) + k_{ins}\}$$

$$arcDel(i,j) = \max_{\forall \text{ arc } b\text{: } \text{dest}(b)=j} \{M(i,\text{ori}(b)) + k_{del}\}$$

$$k_{sub} = \begin{cases} 0 & \text{if S}(a) = \text{S}(b) \\ 0 & \text{if S}(a) \text{ is semivowel or semiconsonant of S}(b) \text{ or vice versa} \\ -\infty & \text{otherwise} \end{cases}$$

3.3 Filtering the candidate detections

All the paths in the dynamic programming matrix that arrive to the last column (j = number of nodes of the graph representing the query) are considered candidate detections. To filter out these detections, algorithm 1 is performed.

Algorithm 1 Algorithm to obtain a filtered list of detections

Require: A list of candidate detections CD
Ensure: A list of filtered detections FD
1: SCD = sort the hypothesis in CD by their score
2: $FD2$ = empty list
3: **while** SCD is not empty **do**
4: h = first element of SCD
5: Move h to $FD2$
6: Delete from SCD all the hypothesis that overlap h
7: **end while**
8: Determine a threshold t considering the score of the elements in $FD2$
9: FD = elements of $FD2$ with a score higher than t
10: **return** FD

The threshold that we have finally used, according to the results obtained in an experimentation using the provided training and development data, is $t = max - 0.5 \cdot sd$, being max the maximum score of the elements in $FD2$ and sd the standard deviation of these scores.

4 The ELiRF-SDTW QbE-STD system

A second system has been developed for the Query-by-Example Spoken Term Detection task. This system is based on the use of the *a posteriori* probabilities of phonemes described in Section 2, Segmental Dynamic Time Warping [8], and the Kullback–Leibler divergence.

4.1 Divergence measure

The Kullback–Leibler divergence (also known as relative entropy) is a measure of how different two probability distributions are. It is defined by:

$$KL(p||q) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)}$$

where p and q are two probability distributions over the same event space $\mathcal{X}$.

To avoid uncertainties, cases where p or q are zeros are treated in a special way. Specifically, we define the following equalities: $0\ log\frac{0}{0} = 0$, $0\ log\frac{0}{q} = 0$, and $p\ log\frac{p}{0} = \infty$

4.2 Segmental Dynamic Time Warping

Segmental DTW (SDTW) [8] is a modification of the well-known Dynamic Time Warping algorithm. The SDTW is to find multiple local alignments of two input utterances. For the QbE task, we search multiple local alignments of the query in the speech data. The main difference between SDTW and DTW is that, while in DTW there is only one start point for the alignment, SDTW permits to start the alignment at any point along the speech data.

Instead of using the DTW typical transitions, we have used (as in other works [1] a different set of transitions (Figure 2) which ensures that the paths found will have a number of frames between half and twice the length of the query.

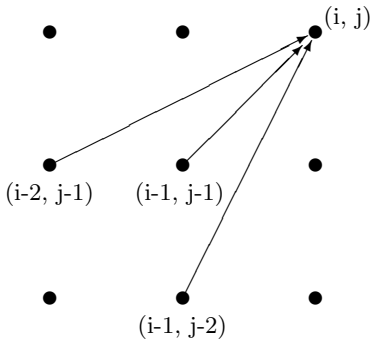


Fig. 2. Transitions set used in the ELiRF-SDTW system

Therefore, considering the two modifications described above, the minimization function at each point is given by:

$$D(i,j) = \begin{cases} 0 & j < 1 \\ \min \begin{pmatrix} D(i-1,j-1) \\ D(i-2,j-1) \\ D(i-1,j-2) \end{pmatrix} + KL(A(i),B(j)) & j \geq 1 \end{cases}$$

where A(i) is the *a posteriori* probabilities of phonemes for the frame i of the speech data, B(j) is the *a posteriori* probabilities of phonemes for the frame j of the query, and KL is the Kullback–Leibler divergence described in the previous subsection.

4.3 Filtering the candidate detections

For each speech data A and each query B a SDTW matrix is obtained. The last row of this matrix ($D(i,|B|)$, $1 \leq i \leq |A|$) contains the scores of all the local alignment candidates. In order to filter this candidate list we have used the algorithm 1 as used in the ELiRF-PhGraph system, but considering the scores normalized by the length of their corresponding paths. The threshold used in algorithm 1 has been experimentally fixed to obtain the best AWVT score for the development set but ensuring at least 10% of coverage.

5 Conclusions

In this work, we have presented two approaches to Query-by-Example Spoken Term Detection. Both approaches are based on a search considering the *a posteriori* phonetic probabilities of both the query and the speech data. In one case the uncertainty is modeled by means of graphs of phonemes, while in the other a modified DTW is used to find the best matching.

Acknowledgments. This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, by the Vic. d'Investigació of the UPV under contract PAID-06-10, and by the Spanish MICINN under FPU Grant AP2010-4193.

References

1. Anguera, X., Macrae, R., Oliver, N.: Partial sequence matching using an unbounded dynamic time warping algorithm. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE (2010) 3582–3585
2. Hazen, T., Shen, W., White, C.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, IEEE (2009) 421–426
3. Zhang, Y., Glass, J.: Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams. In: Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, IEEE (2009) 398–403

4. Akbacak, M., Vergyri, D., Stolcke, A.: Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE (2008) 5240–5243
5. Gómez, J., Sanchis, E., Castro-Bleda, M.: Automatic speech segmentation based on acoustical clustering. Structural, Syntactic, and Statistical Pattern Recognition (2010) 540–548
6. Gómez, J., Castro, M.: Automatic segmentation of speech at the phonetic level. Structural, Syntactic, and Statistical Pattern Recognition (2002) 672–680
7. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Marino, J., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: Third European Conference on Speech Communication and Technology. (1993)
8. Park, A., Glass, J.: Towards unsupervised pattern discovery in speech. In: Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on, IEEE (2005) 53–58

GTTS Systems for the Query-by-Example Spoken Term Detection Task of the Albayzin 2012 Search on Speech Evaluation*

Amparo Varona, Mikel Penagarikano,
Luis Javier Rodríguez-Fuentes, Germán Bordel, and Mireia Diez

GTTS (http://gtts.ehu.es), Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU
Barrio Sarriena, 48940 Leioa, Spain
amparo.varona@ehu.es

Abstract. This paper briefly describes the systems presented by the Working Group on Software Technologies (GTTS)[1] of the University of the Basque Country (UPV/EHU) to the Query-by-Example (QbE) Spoken Term Detection task of the Albayzin 2012 Search on Speech Evaluation. GTTS systems apply the state-of-the-art Brno University of Technology phone decoders for Czech, Hungarian and Russian in two different ways: System A looks for approximate matchings of the best decoding of a spoken query in the phone lattice of the target audio document; System B represents both the query and the audio document in terms of frame-level phone log-likelihoods and scans the audio document for possible matchings, by minimizing the distance between feature vectors and applying heuristic strategies to prune the search space and to validate the hypothesized segments.

Index Terms: Spoken Term Detection, Phone Lattice, String Matching, Phone Log-Likelihoods, Cosine Distance, Heuristic Matching.

1 Introduction

In the Query-by-Example (QbE) Spoken Term Detection task of the Albayzin 2012 Search on Speech Evaluation, the input to the system is an acoustic example per query and hence a prior knowledge of the correct word/phone transcription corresponding to each query is not available [1]. The locations and durations of all the occurrences of spoken queries in the audio documents must be obtained. The task is defined in the same terms as MediaEval 2012 Search on Speech [2].

* This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE11UN065), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds).

[1] http://gtts.ehu.es

2 System A: Exact Matching on Phone Lattices

2.1 System description

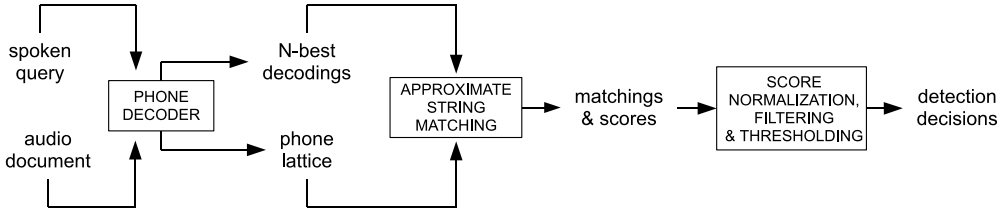


Fig. 1. Processing steps of the GTTS System A for the QbE Spoken Term Detection task of the Albayzin 2012 Search on Speech Evaluation.

Computing phone lattices. As a first step, the open-software Brno University of Technology (BUT) phone decoders for Czech, Hungarian and Russian [3] are applied to decode both the spoken queries and the audio documents. BUT decoders have been trained on 8kHz (SpeechDat) databases, so both the spoken queries and the audio documents are downsampled from 16 kHz to 8kHz.

BUT decoders feature 45, 61 and 52 phonetic units for Czech, Hungarian and Russian, respectively. For each unit, a three-state model is used, so three state posterior probabilities per frame and unit are computed. Since exact (or almost exact) matchings are required to detect queries, the number of phonetic units may be too high for this application. Note that the same sound may be decoded in different ways, in terms of similar (but different) units. To compensate for this effect, the set of units is reduced by defining groups of similar (i.e. highly confusable) units, according to their characterization in the International Phonetic Alphabet (IPA). Besides, the three non-phonetic units used by BUT decoders are fused into a single non-phonetic unit model. Eventually, we use 25 units for Czech, 23 for Hungarian and 21 for Russian.

Let us consider one of the BUT decoders, featuring M phone units, each of them typically represented by means of a left-right model of S states. The posterior probability of each state s ($1 \leq s \leq S$) of each phone model i ($1 \leq i \leq M$) at each frame t, $p_{i,s}(t)$, is directly provided by the phone decoder. When considering a reduced set of units, each unit j clusters a number of similar units, and its posterior probability at each state s and each frame t can be computed by adding the posterior probabilities of all of them:

$$p_{j,s}(t) = \sum_{\forall i \in S_j} p_{i,s}(t) \tag{1}$$

with $1 \leq j \leq R$, R being the number of clusters in the reduced set and S_j the subset of phone units in cluster j. Finally, posterior probabilities are used to produce phone lattices —which encode multiple hypotheses with acoustic likelihoods—, by means of the HTK tool *HVite* [4].

Searching phone lattices. For each spoken query, the phone decoding with the highest likelihood is extracted from the phone lattice by means of the *lattice-tool* of *SRILM* [5]. Only those strings containing more than two phones are considered. Then, the *Lattice2Multigram* (L2M) tool by Dong Wang [6][7][8][2] is applied. L2M takes two inputs: a list of phone strings (the queries) and a list of phone lattices (the documents), and outputs detections in MLF format [4]. The behaviour of L2M is controlled by several parameters, which have been tuned on the development dataset. In particular, *LogLikeliBaumWelch* has been chosen as lattice score computation method and matchings are located under the *ExactMatch* setup (i.e. no edition operations are allowed in matchings).

Handling the scores. Three filters are sequentially applied to MLF detection files:

- *mlf2mlf*: for each detected segment i, a new normalized score is computed in the following way:

$$new_score_i = \log \frac{e^{\frac{score_i}{length_i}}}{\sum_{\forall j \neq i} e^{\frac{score_j}{length_j}}} \quad (2)$$

  where $length_x$ stands for the length (in frames) of a detected segment x, and the summation in the denominator extends over all the detections $j \neq i$ in the audio document.
  Given a set of spoken queries and a set of audio documents, three MLF files are produced, based on the Czech, Hungarian and Russian BUT decoders, respectively. Detection files can be either mixed and processed jointly, or processed independently. In any case, for each audio document, overlapping detections are processed such that only the most likely detection is kept, the remaining ones being discarded.
- *mlf2std*: detection information is converted to the final STD format.
- *std2std*: for each query, only the K most likely detections in all the audio documents are retained, scores are z-normalized and a threshold is applied.

Preliminary Experiments. Table 1 summarizes the results obtained in preliminary experiments on the development set, using different configurations (in all cases, MLF files are mixed and jointly processed, and $K = 50$). The *Actual Term Weighted Value* (ATWV) is used as primary evaluation measure [9], but false alarm and miss error probabilities are shown too. Since best performance was attained when mixing the Hungarian and Russian decoders (HU + RU), this was the setup chosen for the *primary system*. For the contrastive system, we chose the mix of the three decoders (CZ + HU + RU). In both cases, the final threshold was set to 0.20.

[2] http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html

Table 1. Performance of the GTTS System A in preliminary experiments on the development set under different setups.

BUT decoders	Max ATWV	Threshold	P(FA)	P(Miss)
CZ	0.002	0.168	0.00046	0.970
HU	0.004	0.236	0.00056	0.974
RU	0.006	0.222	0.00116	0.947
CZ + HU	0.010	0.125	0.00075	0.956
CZ + RU	0.006	0.256	0.00137	0.935
HU + RU	0.013	0.222	0.00134	0.933
CZ + HU + RU	0.009	0.247	0.00145	0.928

2.2 Training and development data

BUT decoders: training data.

- Czech Decoder (CZ) - 8 kHz, trained on the Czech SpeechDat(E) database, containing 12 hours of speech from 1052 (526 male, 526 female) Czech speakers, recorded over the Czech fixed telephone network.
- Hungarian Decoder (HU) - 8 kHz, trained on the Hungarian SpeechDat(E) database, containing 10 hours of speech from 1000 (511 male, 489 female) Hungarian speakers, recorded over the Hungarian fixed telephone network.
- Russian Decoder (RU) - 8 kHz, trained on the Russian SpeechDat(E) database, containing 18 hours of speech from 2500 (1242 male, 1258 female) Russian speakers, recorded over the Russian fixed telephone network.

Development data. The GTTS System A was developed based exclusively on the materials provided by organizers for this evaluation, consisting of a set of talks extracted from the Spanish MAVIR workshops: 60 spoken queries and 7 audio documents amounting to about 5 hours of speech [1].

3 System B (late submission): Heuristic Matching

3.1 System Description

The contrastive (late) GTTS system uses a frame-level sequence of phone log-likelihoods to represent both the query and the audio document. The BUT decoders for Czech, Hungarian and Russian are applied to downsampled (8 kHz) signals to get frame-level phone log-likelihoods, at a rate of 100 frames per second. Phone log-likelihoods are stacked in a single feature vector, those corresponding to non-speech units being left out. Thus, feature vectors include 42 log-likelihoods from Czech, 58 from Hungarian and 49 from Russian, which amounts to 149 dimensions.

Based on the above described representation, multiple occurrences of the query inside the audio document could be found just by defining a suitable distance measure between two feature vectors and applying a Dynamic Time

Warping (DTW) approach which minimizes the accumulated distance. The time and space complexities of this DTW-based approach would be in $\Theta(m \cdot n)$, where m: length of the query and n: length of the audio document, with $m \ll n$. We tried DTW using different distances (Euclidean, Mahalanobis, cosine), with and without z-normalization, but got unsatisfactory results on a set of toy examples. Then we tried a slightly different approach, based on a frame-by-frame greedy search for matchings (that is, taking locally optimal decisions) and on some heuristic pruning and validation criteria. Though space and time complexities were still in $\Theta(m \cdot n)$, this *heuristic matching* approach yielded much better results than DTW on the toy examples used for development.

Feature normalization. Let us consider a spoken query Q, represented by the sequence of D-dimensional feature vectors $Q = \{q_1, q_2, \ldots, q_m\}$, and the audio document $A = \{a_1, a_2, \ldots, a_n\}$ on which the search is performed. First, we *z-normalize* the feature vectors of both Q and A according to the means $\mu = [\mu_1, \mu_2, \ldots, \mu_D]^t$ and standard deviations $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_D]^t$ estimated on the audio document A. Given a feature vector v, the normalized vector $\hat{v}$ is given by:

$$\hat{v} = \left[\frac{v_1 - \mu_1}{\sigma_1}, \frac{v_2 - \mu_2}{\sigma_2}, \ldots, \frac{v_D - \mu_D}{\sigma_D} \right]^t \tag{3}$$

Distance measure. We implemented three distance measures to be applied on the z-normalized features: Euclidean, Mahalanobis and cosine. Best results in preliminary experiments on a set of toy examples were obtained with the cosine distance, defined as follows:

$$d(v, w) = \frac{\sum_{i=1}^{D} \exp(v_i + w_i)}{\sqrt{\sum_{i=1}^{D} \exp(v_i + v_i)} \sqrt{\sum_{i=1}^{D} \exp(w_i + w_i)}} \tag{4}$$

Note that dot products are performed in the space of likelihoods, so the z-normalized log-likelihoods used as features are added and the result is exponentiated and accumulated.

Within-query maximum distances. In order to define query-dependent thresholds, for each query Q the maximum distance between two feature vectors being k frames away from each other, q_i and q_{i+k}, is computed and stored, for $k = 1, 2, \ldots, 10$:

$$dmax(Q, k) = \max_{i=1,\ldots,m-k} d(q_i, q_{i+k}) \tag{5}$$

Search procedure. The search procedure consists of a frame-by-frame sequence of matching attempts, so that up to n different matchings are tried in the worst case. To avoid repeated distance computations, the whole matrix of distances between all the pairs of feature vectors (q_i, a_j), q_i from the query Q

($i \in [1, m]$) and a_j from the audio document A ($j \in [1, n]$), is computed in first place, so time and space complexities are both in $\Theta(m \cdot n)$.

The length of segments matching the query is heuristically bounded by two warping factors, w_{min} and w_{max}. Segments with less than $w_{min} \cdot m$ frames are discarded. On the other hand, if a partial match starting at frame i involves more than $w_{max} \cdot m$ frames, the current search is abandoned and a new search is started at frame $i+1$ (see Figure 2). In this work, $w_{min} = 0.5$ and $w_{max} = 2.0$.

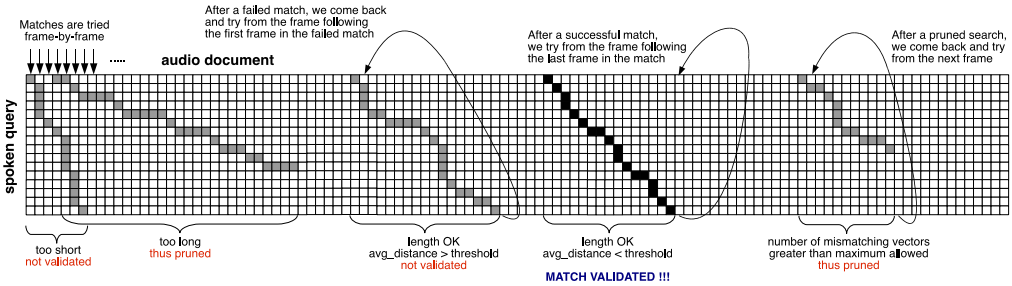


Fig. 2. Heuristic matching consists of a frame-by-frame sequence of matching attempts. Decisions are taken in a greedy fashion, minimizing the distance between the aligned vectors. Several heuristically fixed pruning and validation thresholds are applied.

A threshold θ was established so that if the distance between two feature vectors was greater than θ, we considered it a *mismatch*. We defined θ as a linear combination of within-query maximum distances, as follows:

$$\theta = \theta(Q,\alpha) = (1-\alpha) \cdot dmax(Q,1) + \alpha \cdot dmax(Q,10) \tag{6}$$

After tuning on some toy examples, best performance was found for $\alpha = 0.30$.

A maximum number of mismatches L was allowed during search, so that, given a partial match starting at frame i, if the number of mismatches exceeded L, the current search was abandoned and a new search was started at frame $i+1$ (see Figure 2). In this work, L was heuristically fixed to a fraction of the length of the query: $L = \lambda \cdot m$, with $\lambda = 0.20$.

For each segment $A[i,j]$ matching a query Q (and fulfilling the above described conditions), a threshold δ was applied, so that only if the average distance between feature vectors of Q and $A[i,j]$, $d_{avg}(Q, A[i,j])$, was lower than δ, the segment was accepted and $d_{avg}(Q, A[i,j])^{-1}$ was output as score. The threshold δ was also defined as a linear combination of within-query maximum distances, as follows:

$$\delta = \delta(Q,\beta) = (1-\beta) \cdot dmax(Q,1) + \beta \cdot dmax(Q,2) \tag{7}$$

After tuning on some toy examples, best performance was found for $\beta = 0.40$.

Preliminary Experiments. Using parameter values optimized on a set of toy examples ($w_{min} = 0.5$, $w_{max} = 2.0$, $\alpha = 0.30$, $\lambda = 0.20$ and $\beta = 0.40$), the best performance in preliminary experiments on the development set was found when the score threshold was set to 8.0.

3.2 Training and development data

As for System A, BUT decoders were applied to queries and audio documents to get frame-level phone log-likelihood feature vectors. Development data were also the same used for System A. Details have been already provided in Section 2.2.

References

1. J. Tejedor, D. T. Toledano, and J. Colas, "The Albayzin 2012 Search on Speech Evaluation," in *VIII Jornadas en Tecnologa del Habla and III Iberian SLTech Workshop, IberSPEECH*, Madrid, Spain, 2012.
2. F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, and N. Rajput, "The Spoken Web Search task," in *MediaEval 2012 Workshop*, Pisa, Italy, 2012.
3. P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, FIT, BUT, Brno, Czech Republic, 2008.
4. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (Version 3.4)*, Cambridge, 2006.
5. A. Stolcke, "SRILM - an extensible language modeling toolkit," in *ICSLP*, 2002, pp. 257–286.
6. D. Wang, S. King, and J. Frankel, "Stochastic pronunciation modelling for out-of-vocabulary spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.
7. D. Wang, J. Tejedor, S. King, and J. Frankel, "Term-dependent confidence normalization for out-of-vocabulary spoken term detection," *Journal of Computer Science and Technology*, 2012.
8. D. Wang, S. King, and J. Frankel, "Direct posterior confidence estimation for spoken term detection," *ACM Transactions on Information Systems*, 2012.
9. J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," in *ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, Amsterdam, 2007.

Telefonica Research system for the Query-by-example task at Albayzin 2012

Xavier Anguera

Telefonica Research,
Edificio Telefonica-Diagonal 00, 08019, Barcelona, Spain,
xanguera@tid.es

Abstract. In this paper we describe the Telefonica Research submission to the Albayzin 2012 Evaluation. In particular, we participated in the query-by-example task within the search on speech evaluation. Our system is a a zero-resources approach by using a variant of the Dynamic Time Warping algorithm. We also preprocess and post-process the signal to make the features as much speaker independent as possible, to eliminate silence frames usually causing an increase in false alarms and eliminating unnecessary overlapping matching segments. The results are low due to the difficulty of the task and the strictness of the used metric, but are representative of the potential of these techniques for search on speech when no information is available a priori on what language or acoustic conditions we are facing.

Keywords: Query by example, pattern matching, low resources

1 Introduction

The query-by-example subtask proposed within the search on speech task in the Albayzin 2012 evaluation proposes to search for occurrences of a given query within a spoken database where the query is given in acoustic form and where no transcription is available neither of the query nor of the spoken database. This task, together with the spoken term detection task where a textual query needs to be searched over the audio corpus, are gaining importance in the last few years and are being applied to cases where not much data is available a priori to train the systems being used. This is the case of languages with few available resources or when encountering very adverse acoustic conditions for which no pretrained models are available. For this talk two kinds of algorithms are usually applied. On the one hand, researchers are looking again into what was being used before the arrival of Hidden Markov Models and improving them to be adapted to the new challenges ahead. This includes algorithms that search for patterns in the speech, like Dynamic Time Warping (DTW).

One of the first algorithms to improve on the classical DTW algorithm was proposed in [8] where an iterative search was done in an audio signal looking for repeating patterns. Later on, other researchers have proposed alternative solutions [11, 10, 9] that allow systems to be more accurate and/or fast. Recently, in

[7] we have seen that this task does not require any more of expensive equipment to run and is very scalable to search on high volumes of data.

An alternative to DTW-like algorithms are the standard phonetic search approaches. If the language spoken in the recordings is known, does not change over time and we have training data for it, very good results can be obtained (for reference, see any of the literature in the spoken term detection community). Problems with these algorithms often come when the language is unknown or we have a low-resources situation, in which similar languages have to be adapted to be able to represent the data being searched, with clear accuracy problems. On their favor, these systems are born scalable as they can use lots of the speech recognition speedup methods that have been proposed over the years, although their computational requirements usually involve heavy machinery (maybe with parallel processors) to be able to process the data as fast as possible. In 2011, the Mediaeval SWS evaluation [5] proposed a query-by-example task on indian data and in [6] results are given comparing both phonetic-based and pattern-matching based approaches.

The approach we applied for this evaluation is a zero-resources approach by using a variation of the DTW algorithm called Segmental-DTW. Such algorithm is able to search for a given query throughout the reference data by allowing a certain time warping between the query and the reference instances of the queried word. In this paper we describe the system implemented and the results we obtained in the evaluation. In general, given the difficulty of the task, we are happy with the results, although there are some possible areas of improvement as we will highlight in the conclusions and future work section.

2 System description

For the query-by-example evaluation we have applied a zero-resources matching approach based on the well-known Dynamic Time Warping (DTW) algorithm. The whole algorithm (from signal input to results output) is depicted in figure 1. First, acoustic features need to be extracted from the signal. For this we first obtain MFCC-39 coefficients from the signal (13 Cepstra + 13 delta + 13 double delta). As shown in [2] the MFCC features themselves are not sufficiently speaker independent. For this reason we use them to compute posteriorgram features by using a posteriors background model we describe in more detail in Section 2.1 below. Once posteriors are computed we trim out those frames that are classified as silence or very low speech as described in Section 2.2. The resulting features are stored on the database for reference features and sent to the matching algorithm for the query features. Section 3 describes in detail the matching algorithm we used for this evaluation. After the matching algorithm returns all possible matching paths and their scores, we postprocess the results to merge all those matches that overlap with each other, as explained in Section 2.3.

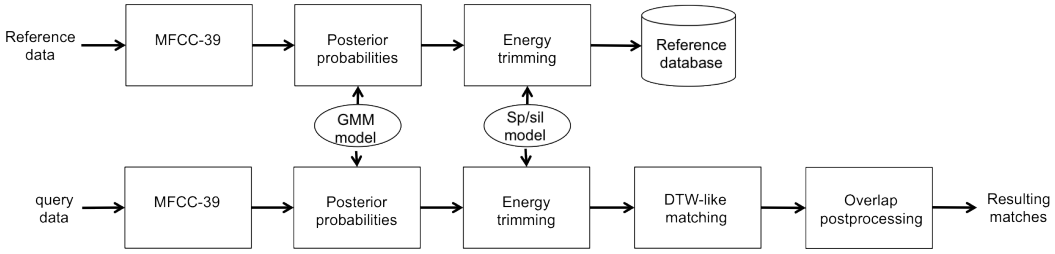


Fig. 1. *General system blocks for QbE task*

2.1 Posteriors Background Model

Posterior probabilities have been successfully used in pattern matching for some time [3]. Several methods have been proposed to obtain the posterior probabilities. In our systems we use Gaussian posteriors obtained from a GMM model that has been previously trained on all available reference data (i.e. development and testing data). The GMM is trained using a combination of EM and K-means iterations in order to maximize the discovery and separation of automatically discovered acoustic regions in the acoustic space. For more information on the model refer to [1].

2.2 Speech/silence labeling

One of the biggest enemies of pattern matching approaches is silence, as silence usually matches very well with silence, thus returning many false alarms unless it can be trimmed back. To eliminate silence from our input data (both queries and references) this year we trained a speech/silence classifier using GMM models in a non-supervised way. First, we gather the 10% of acoustic frames with lowest energy from our training data (the reference files in the development set). With these frames we train a one Gaussian silence model and with the rest we train a 4-Gaussian speech model. Then we iteratively assign each frame in the training set to the closest model and retrain the models. This usually increases the number of frames in the silence model. We stop after 20 iterations or when the difference in number of frames between two consecutive iterations is very small. We store the Gaussians in the speech model ordered by their mean energy. In order to label the data using this model we no only assign each frame to the most likely model, but in the case of speech, we record which of the Gaussian mixtures is the closest one. In test we consider as silence (and therefore do not use for matching) any frame that is labelled as silence or is assigned to the lowest speech Gaussian.

2.3 Matching segments overlap detection

The overlap detection module is used on all matching segments returned by the matching module to reduce the number of segments finally returned as matches.

It is common that matching algorithms based on pattern matching return a high number of possible paths around the true matching path, as frames around the optimum path also contain very low distances. In our implementation we allow the matching algorithm to register all possible paths it finds and then we search for those paths that have an overlap in the reference higher than 50% (note that the query segment is always the same, covering all the query time). In our implementation we payed special attention for matching paths that are contained within bigger paths. These are sometimes specially good subsequences within a longer sequence. Given two paths that are in overlap, in the current system we select that one path that had a higher average score (i.e. in average it matches between the query). In the past we would merge the paths by selecting the smallest and biggest of the start and end points, respectively. We are not following this approach anymore as the matching algorithm requires for some precision in the detection of the matching segment, therefore it is better to end up with shorter paths than with much longer paths.

3 Segmental Dynamic Time Warping

Given two sequences, X and Y of posterior probabilities, respectively obtained from the query and the reference audio recordings, we compare them using a DTW-like algorithm. The standard DTW algorithm returns the optimum alignment between any two sequences by finding the optimum path between their start $(0,0)$ and end (x_{end}, y_{end}) points. In our case we constraint the query signal to match between start and end, but we allow the phone recording to start its alignment at any position $(0, y)$ and finish its alignment in whenever the dynamic programming algorithm reaches $x = x_{end}$. This algorithm is generally called Segmental Dynamic Time Warping (SDTW). Although we do not set any global constraints in the algorithm (in order to allow any matches to exist in any position in the reference recording), the local constraints used by the dynamic programming are set so that at maximum 2-times or $\frac{1}{2}$-times warping is allowed by choosing the path that minimizes the cost to reach position (i, j) as

$$\text{cost}(i,j) = (d(i,j) + \min \begin{cases} D(i-2,j-1))/(\#(i-2,j-1)+3) \\ D(i-2,j-2))/(\#(i-2,j-2)+4) \\ D(i-1,j-2))/(\#(i-1,j-2)+3) \end{cases} \quad (1)$$

Where $D(i,j)$ is the accumulated (non-normalized) distance of all optimum paths until position (i,j), $d(i,j)$ is the local distance between frames x_i and y_j from both compared sequences, and $\#(i,j)$ is the number of jumps of the optimum path until that point. Note than when normalizing the different possible paths we slightly favor the diagonal match by considering the Manhattan distance between the current point and the previous point. Such inequality is also kept when updating the number of jumps matrix, thus favoring paths that contain many diagonal matches. In the current implementation we run the algorithm independently for each reference file and each query file, but keeping all

of them in memory for fast access to the data. Once the cost matrix has been computed for a given query and reference file, we find in the last row of such matrix (positions (N, j) where N is the last query frame and j are all possible reference frames) the aggregate scores of the best paths matching query and reference data. By locating the minima of this function and applying a threshold we can find the hypothetical matches between our query and the reference. By performing a backtracking of the selected matches we can find where in the reference data the match started. Such position will always be between 2-times and $\frac{1}{2}$-times the length of the query signal, as imposed by the local constraints imposed above.

4 Evaluation Database and Results

As described in [4], data provided by the organizers for system training, development and evaluation belong to recordings of the MAVIR workshops in 2006, 2007 and 2008 (Corpus MAVIR 2006, 2007 and 2008) corresponding to Spanish language. Training/development data amounts about 5 hours of speech material in total. For evaluation, test speech data amounts about 2 hours in total. Queries are acoustic snippets of the data containing words spoken several times in the query and reference files. In general we observed that the query snippets are quite short (less than 1 second) and contain usually just one word per query with no silences at the beginning or end. These queries have been obtained from the reference data by cutting a particular instance of the word, usually uttered within a sentence with other words in coarticulation just before and after it. Although this is not the optimal use case for a query-by-example application, it is a challenging one that systems should be able to handle.

The metrics used are the same as used in the spoken term detection task originally developed by NIST in the STD 2006 evaluation and described also in [4]. There are two main metrics. On the one hand, the Actual Term Weighted Value (ATWV) computes an individual weighted sum of misses and false alarms for each query term and given a working general threshold, and then obtains the average value. On the other hand, the Maximum Term Weighted Value (MTWV) looks at all possible thresholds and selects the optimum one. It is desired that the threshold defined on the development data also be optimal for the evaluation, so that the system can generalize to unseen datasets.

Table 1 shows the official results we obtained in the evaluation for the development and the evaluation sets. The threshold used in the evaluation set is the same one we optimized for the development set. We noticed how our system obtains quite low scores for the MTWV, which is due to the high importance that the scoring gives to the false alarms. On the development set, by the ATWV system, we obtain 45 correct matches, 33 false alarms and 982 misses. It is clear that the false alarms are driving the scoring, which is one of our main current problems.

The relationship between misses and false alarms can be seen more clearly in Figure 2, which shows the DET curve of misses versus false alarms for the

Table 1. Official Evaluation Results

Metric	dev set	eval set
MTWV	0.0455	tbd
ATWV	0.0425	tbd

development set. We can see how our system gives results far better than random, although it still have a long way to go to call this task as solved.

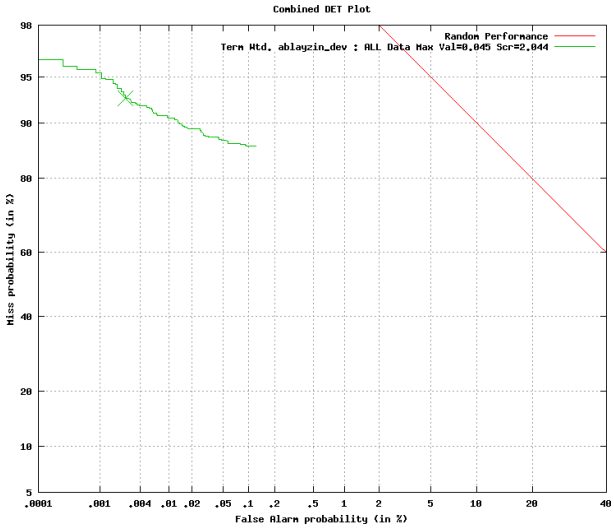


Fig. 2. *DET curve for the development set*

5 Conclusions and Future Work

In this paper we present the Telefonica Research system we presented to the Albayzin 2012 query-by-example task. For this we used a zero-resources approach based on the segmental-DTW algorithm. Our results are far from optimal and indicate how difficult the task is. We are planning on working out some algorithms to reduce the number of false alarms, and therefore should obtain better MTWV in the future.

References

1. X. Anguera: Speaker Independent Discriminant Feature Extraction for Acoustic Pattern-Matching. In Proc. ICASSP 2012, Kyoto, Japan.
2. A. Muscariello, G. Gravier and F. Bimbot: Zero-resource audio-only spoken term detection based on a combination of template matching techniques In Proc. INTERSPEECH 2012, Florence, Italy.

3. Y. Zhang and J. Glass: Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams In Proc. ASRU 2009, Merano, Italy.
4. J. Tejedor, D. T. Toledano and J. Colas: The ALBAYZIN 2012 Search on Speech Evaluation In Proc. Iberspeech 2012, Madrid, Spain.
5. MediaEval 2011 Workshop, Pisa, Italy, Sept. 2011.
6. F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier6, C. Heerden, G.V. Mantena, A. Muscariello, K. Prahallad, I. Szoke, and J.Tejedor: The spoken web search task at mediaeval 2011 In Proc ICASSP 2011, Kyoto, Japan.
7. A. Jansen, B. V. Durme: Indexing Raw Acoustic Features for Scalable Zero Resource Search In Proc. Interspeech 2012, Portland, OR, USA
8. A. Park and J. Glass: Unsupervised Pattern Discovery in Speech In IEEE TASLP, vol. 16, no. 1, January 2008
9. A. Jansen, K. Church and H. Hermansky: Towards Spoken Term Discovery at Scale with Zero Resources In Proc. Interspeech 2010, Tokyo, Japan.
10. A. Muscariello, G. Gravier and F. Bimbot: Variability tolerant audio motif discovery In Proc. Int. Conf. on Multimedia Modeling, 2009
11. X. Anguera, R. Macrae and N. Oliver Partial Sequence Matching using an Unbounded Dynamic Time Warping Algorithm In Proc. ICASSP 2010, Dallas, TX, USA

Albayzin 2012 search on speech @ ViVoLab UZ

Antonio Miguel, Jesús Villalba, Alfonso Ortega and Eduardo Lleida

Voice Input Voice Output Laboratory (ViVoLab)
Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{amiguel,villalba,ortega,lleida}@unizar.es

Abstract. This paper describes the system developed for the Albayzin 2012 spoken term detection and keyword spotting Evaluation. The main feature of the system is the fusion of the scores provided by three recognizers/subsystems. To provide those scores we have built three main acoustic-phonetic recognizers. The first one is a continuous density HMM with triphone contextual units. The second one is a neural network multilayer perceptron hybrid HMM with biphone contextual units. The third one is a continuous density HMM with monophone incontextual units whose Gaussians are factor analyzed at the sentence level. For each of those recognizers, two scores are obtained. The first one is the edit distance using the phoneme output sequence and the second one is the sum of the aligned confidences obtained from the recognition lattice. Finally, scores are calibrated with Bosaris toolkit.

Keywords: Spoken term detection, HMM, MLP, Factor analysis

1 Introduction

The system proposed by the Vivolab-UZ team from I3A at University of Zaragoza is the fusion of three recognizers or subsystems which are explained later. The first recognizer is a standard triphone contextual HMMs with mixtures of diagonal covariance Gaussians as observation density functions in the states. The second one is a biphone contextual HMM where the observation probabilities are obtained from a multilayer perceptron, MLP. The third one is a monophone incontextual HMM where the observation probabilities are obtained from mixtures of full covariance factor analyzed Gaussians, where the factor analysis is done at the sentence level.

This paper is organized as follows, in Section 2, the system is presented and each of the different subsystems discussed. The fusion system and the training and development data are also explained in Section 2. Conclusions are shown in Section 3.

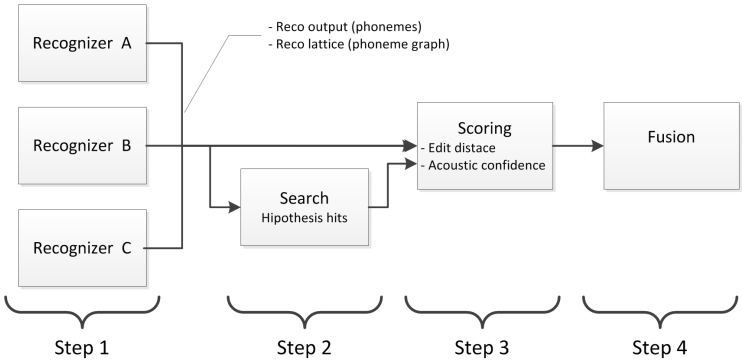


Fig. 1. Modules and processing steps of the search on speech system

2 System Vivolab-I3A-UZ

2.1 System description

In Figure 1 it is displayed the block diagram of the whole system. We can see how the three subsystems are used to produce the final scores. The process consists mainly in four steps.

The first step is the recognition of the sequence of phonemes and generation of lattices, which will be used by the other parts of the system for the search process. For this purpose we have used the tool HDecode from the HTK well known software [1] for all the systems. The particular details of each of the three proposed recognizers are explained later.

The second step is the generation of hit hypothesis. The idea in the step is to use the recognizer outputs offline at search time to produce the hit hypothesis, which we will refer to as trials. The threshold for generating these hypotheses is quite low in order to miss a small number of targets even there are many non-target trials.

The third step is the scoring, we use the previously generated search hits with each recognizer to provide the scores to the fusion system. We obtain two scores for each recognizer: the edit distance and an acoustic confidence measure obtained from the lattice.

The final step is to merge the previous scores and fuse them to obtain a single score, which will be used to classify as target and non-target. For the fusion we have used a linear model trained with the Bosaris toolkit [2].

Step 1, Recognizer A The first of the recognizers is the closer of the three to a standard phonetic decoder, since it uses a trigram language model and contextual triphone three state HMM units with mixtures of Gaussians in the

[1] http://htk.eng.cam.ac.uk/
[2] https://sites.google.com/site/bosaristoolkit/

states (diagonal covariance matrices). The features used in this recognizer are the ETSI MFCCs with derivatives and cepstral mean compensation, with a total dimension of 39.

Language model

In the case of a phonetic decoder, the output of the recognizer is the sequence of phonemes, then the phoneme is acting as a word. The trigram language model provides the likelihood of sequences of three phonemes $p(ph1|ph2, ph3)$, in the same way as if they were the words of a large vocabulary decoder. We have trained the phoneme trigram language model using the phonetic transcription of the training set, which is described later in this section. The language model has been estimated using the SRI Language model toolkit [3] with the default configuration.

Acoustic model

The acoustic model, as described before, is composed of standard continuous density contextual triphone HMMs trained using the audio described later in this section. The number of components in the mixtures is 32 per state and we use three states in the contextual units. Nevertheless, we included a modification in the standard procedure of HTK to train this kind of models. The acoustic units in our case were more specific and we defined special phoneme units for beginning and ending phonemes of each word, in a similar idea of head and tail models found in the literature [1], but in our case is used to obtain more specific phoneme models, not digit digit word models. We provide an example of how the dictionary for training these units is written for some random words:

```
creido =k r e j D o=
llegaban =L e G a B a n=
enfrenten =e m f r e n t e n=
milagro =m i l a G r o=
organizamos =o r G a n i T a m o s=
flagrante =f l a G r a n t e=
aseguraba =a s e G u r a B a=
...
```

(Example of dictionary using specific phonemes)

We use the character '=' to specialize these phonemes being head or tail models. After this definition is done, we proceed to train the contextual units and to tie the more similar using the same procedure as the standard contextual training [2], so that the final number of units depends on the seen examples and the unseen are tied to those seen using a phonetic clustering tree. With this method the number of units grows with respect to a standard model but it still can be managed thanks to the clustering.

Step 1, Recognizer B The second recognizer is an hybrid MLP HMM recognizer. The basic HMM unit is the biphone and we have adopted a simple

[3] http://www.speech.sri.com/projects/srilm/

solution to provide the emission probabilities. The language model in this case is obtained as the previous one using the SRILM toolkit, but in this case the phonemes are not extended to the head tail set.

MLP

There are several examples of hybrid recognizers in the literature [3]. There have been also approaches to context dependent hybrid systems [4, 5] also applied to large vocabulary systems [6]. Out approach has two slight differences with previous approaches: the feature extraction and the fusion of classifiers.

The first change is the feature vector which is fed into the MLP. In our experiments we have concatenated two kinds of features: the standard MFCCs and the log filter bank outputs in a vector stacking several consecutive frames. A total of 110ms of speech (11 frames are merged at each vector).

The second modification is the fusion of several classifiers which we explain later in detail, some of them are classifying articulatory classes. This approach is recently being used in several works [7–9].

The MLP that has been built for this phonetic recognizer is a simple approach to a larger number of layers, which is recently providing very good results [10, 11]. As a simpler method to obtain a classifier with more hidden layers, we have built intermediate MLPs to classify some different objective classes like speech non speech, articulatory and phonemes. Then we have used these outputs as the input of a fusion MLP which is the final phoneme class. This process is explained in the Figure 2, where we can see the two stages of MLPs applied to classify the audio stream

MLPs stage 1

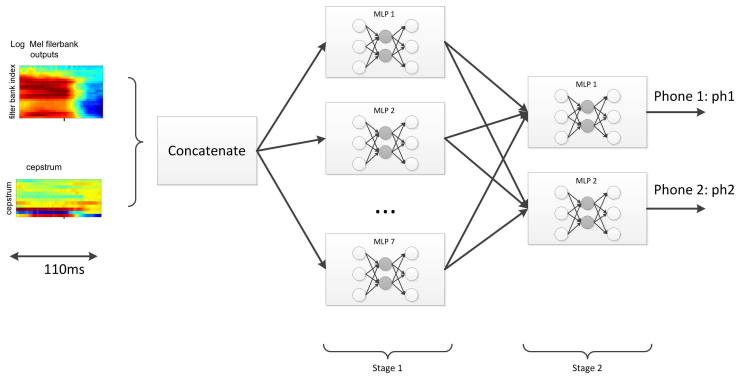


Fig. 2. Recognizer B, MLP block diagram

We now describe the first stage of mlps We have trained 7 networks from the raw audio data with different class targets. The input is built by concatenating MFCC static frames and log filter bank output frames, we will refer to them as FB features. The 12 static MFCCs and the log energy have been extracted using the standard ETSI [12], with a windows size of 25ms and advance of 10ms. The

26 FBs plus the log energy were extracted using HTK with a windows size of 25ms and advance of 5ms, since we found in preliminary tests that this smaller advace step helped the classification accuracy. For the MFCCs we took 11 frames and from the FBs we used 21 frames since the window is smaller. The dimension of the input vector is then: $D = 13 \times 11 + 27 \times 21 = 710$.

In order to obtain frame by frame labels we used a standard HMM alignment similar to the one in the previous system: contextual triphones and mixtures of Gaussians. The seven networks trained from these source features are the following:

- MLP 1: **Speech/Non-Speech class**. The first network is trained using as target classes only speech and non speech for the silence models. In our case all the frames aligned to: 'sil' and 'sp' are the non-speech model and the rest are associated to the speech model.
- MLP 2: **Central articulatory class**. We have used the aligned phonemes (central) to obtain articulatory classes for the following classes: 'Stop', 'Nasal', 'Fricative', 'Liquid', 'Flap', 'Vowel', 'Glide', 'Silence', 'Palatal'.
- MLP 3: **Left articulatory class**. The same set of classes from the previous type but we use the phoneme of the left context to generate it.
- MLP 4: **Right articulatory class**. We use the phoneme of the right context.
- MLP 5: **Central phoneme class**. The class is defined by the phoneme (central) given by the alignment.
- MLP 6: **Left phoneme class**. The class is defined by the left context phoneme given by the alignment.
- MLP 7: **Right phoneme class**. The class is defined by the right context phoneme.

Table 1. Sizes of the MLPs for the first stage, where the hidden layer sizes are also displayed in a vector

MLP	name	D	H	C
1	Speech/Non-Speech	710	[1000,1000]	2
2	Central articulatory	710	[1000,1000]	9
3	Left articulatory	710	[1000,1000]	9
4	Right articulatory	710	[1000,1000]	9
5	Central ph	710	[1500,1500]	27
6	Left ph	710	[1500,1500]	27
7	Right ph	710	[1500,1500]	27

The sizes of the networks are detailed in the Table 1, where D is the input dimension, H the sizes of the hidden layers and C the number of output classes.

As it has bee done recently [13, 14], in order to efficiently train all those networks for the database described later, we have developed a training tool which makes use of a graphic processing unit GPU instead of the CPU for the heavier calculations. This procedure can accelerate the training substantially. For our experiments we used two types of GPUs: Tesla M2090 and GeForce GTX 560.

MLPs stage 2

At stage two we use the previous networks as features for a network which will be trained for phoneme outputs. The input dimension is the sum of the outputs of the previous networks: $D = 2 + 3 \times 9 + 3 \times 27 = 110$.

Table 2. Sizes of the MLPs for the second stage

MLP	name	D	H	C
8	Central ph	110	[1000]	27
9	Right ph	110	[1000]	27

We train two of these fusion networks, the first one is trained for the central phoneme classes and the second one is trained for the right context phoneme. The sizes of the fusion networks are shown in Table 2.

Finally the probability of a contextual phoneme is needed in the search while decoding. We have defined a simple probability for the biphone context as factor of independent probabilities of the central and right phonemes defined previously at stage 2, and this result is used in the decoding.

$$p(\text{ph1+ph2}) \simeq p(ph1|MLP8)p(ph2|MLP9) \quad (1)$$

Step 1, Recognizer C The third recognizer is the simpler from the point of view of number of units and Gaussians. Nevertheless we have adopted a powerful solution in this recognizer to reduce the impact of speaker and channel variability as possible. The solution is based on the eigenvoice models for rapid speaker adapting [15]. The dimension of the features in this case is 39, since we are using the ETSI MFCCs with derivatives and cepstral mean compensation as in recognizer A.

The approach is to train the factor loading matrices in the same way as it is done for factor analysis in current speaker models [16], which means obtaining a single vector of factors (ivector) for each sentence. This can be done in training, since each sentence only has one speaker. In the testing of Mavir we have to adopt a simple solution if we do not use a segmentation system, since there are multiple speakers in the same file. The solution that we have used for the Mavir corpus is to split the files in sections of 10s where we obtain the factor vector.

The procedure for training will be the following:

- First, a baseline 16 Gaussian (diagonal covariance) mixture monophone model is learnt as base.
- Then we use the previous mixtures as base for a mixture of factor analyzers (FAs) [17] with an inner dimension of 20 and we iterate this models.
- The model obtained is used as base to obtain the adapted models to each of the speakers in training. Then we perform a SVD analysis to obtain the eigenvoices using the means of all the models concatenated in a supervector. The inner dimension of the analysis was set to 50. This is sometimes part of the procedure used in speaker factor analyzed models to obtain an initial model [18].
- Finally we iterate the factor analyzed HMM by extracting the Viterbi segmentation using a non factored model (the standard HMM with mixtures of Gaussians) and then applying the ivector extraction as in speaker FA, also called vector of factors.
- When we have the ivector, we update all the sufficient statistics to update the factor analyzed Gaussians and perform the M step.

The procedure at decoding time will be the following

- Fist we use the previously trained model and we estimate the ivector in the same way as in the training. Nevertheless, we have to provide a transcription for the Viterbi alignment, then we use the standard recognizer output as the transcription.
- Using this model we configure the new model means and covariances to obtain a speaker adapted model [15, 19].
- Finally the adapted model is used to provide the transcription and lattices for the scoring system.

Step 2, Generation of hypothesis hits For the generation of hypothesis we used the output phonemes decoded by the previous recognizers as the source text for an edit distance search. Each possible target will be any substring which has an edit distance with the words we are searching of less than 50% of the length of the word in phonemes. This means that we need to correctly decode half of the phonemes of a word to be a hypothesis hit. Then we take all the hits provided by the different recognizers and we merge them when there is overlap using the best one. This will provide a list of trials, which will be scored in the next module. A trial will include the following information: the starting time, the ending time and the word. In the case of training and development data the trial has also information about the true labels whether is target or not target. This information will be used in the fusion step to train the system.

Step 3, Scoring Once the list of trials has been created, we take each of them and we evaluate two scores for each recognizer, the edit distance using the recognition output and a confidence measure based on the lattice.

Edit distance

The edit distance is a well known algorithm where we measure the cost in number of substitutions, deletions and insertions to transform one string into another. For each trial we take the times stored in the previous step and measure the edit distance of the recognized phoneme sequence which lays between the limits and the word phoneme sequence. In order to use this as score, we simply use the negative of the edit distance divided by the length of the word as score for each trial.

Confidence measure

To obtain the acoustic confidence measure we follow these steps:

- First, we determinize the lattice so that a smaller and more useful graph is used in the next step. For this task we use the HTK command HLRescore.
- Second, we obtain an acoustic mesh graph of the decoded phoneme lattice using the lattice-tool from the SRILM toolkit.
- Third, The confidence calculated in the acoustic mesh graph is used in a modified edit distance algorithm where instead of costs all equal to one, we only sum the confidence of matching phonemes with the search word.

Then the score of a trial is the sum of the confidences through the acoustic mesh of the searched word between the time limits defined by the trial. This score is also normalized by the length of the word.

Step 4, Fusion We take all the scores for each trial and compose a matrix with the information of target and non target. This information is used in the Bosaris toolkit to learn a linear fusion. For the costs of miss and false alarm we have used the values: 1 and 0.1. For the probability of target prior we have taken: $10-2$, which is estimated from our target rate in the hypotesis generator. Nevertheless, we have to say that this method is not providing the true error of the task since we have missed many true hypothesis hits in the previous step, which are not accounted as errors now. Then the error measure has to be calculated on the final result of the system.

2.2 Train and development data

Training acoustic models For training the acoustic models the following corpora have been used:

- Albayzin [20]. This is a phonetic balanced corpus which was recorded in a noise free environment. We have used 13600 files with a total length of 12.80h.
- Speech-Dat-Car [21]. This is a corpus recorded in a car in different driving conditions, where we have only taken the close-talk microphone. The number of files that we have used is 25378 and the total length 18.85h.
- Domolab [22]. This corpus was recorded in a domotic environment. We have used the close talk microphone and one of the lapel microphones. A total of 11998 files and 9.33h length.

– TCSTAR [23]. The TCSTAR corpus consists on a series of transcriptions of Spanish parliament sessions. We have used in the training a total of 43357 audio sections with a length of 111.89h.

A total length of 152.87h of training for the acoustic models.

Training and development of fusion For training the fusion we have used the previously explained procedure to obtain a list of trials. To obtain this list we have used all the Mavir [4] files provided for the training.

3 Conclusions

We have presented the fusion of three subsystems for the spoken term search and keyword spotting evaluation. The methodology has been oriented to obtain a series of hypothesis hits using the edit distance with a loose threshold. These hits are fed to the scoring and fusion system in order to rank them according to the knowledge provided by more sources than a single system. The final scores have been calibrated using the training set, so that in test time the found words can be written to the output files when the fused score is greater than the calibrated threshold.

Acknowledgments. This work was funded by the Spanish Ministry of Science and Innovation under project TIN2011-28169-C05-02.

References

1. Juang, B.h., Chou, W., Lee, C.h.: Minimum Classification Error Rate Methods for Speech Recognition. IEEE Transactions on Speech and Audio Processing **5**(3) (1997) 257–265
2. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. Proceedings of the workshop on Human Language Technology - HLT '94 (1994) 307
3. Morgan, N., Bourlard, H.: An Introduction to Hybrid HMM/Connectionist Continuous Speech Recognition. IEEE Signal Processing Magazine (1995) 25 – 42
4. Bourlard, H., Morgan, N., Wooters, C., Renals, S.: CDNN: a context dependent neural network for continuous speech recognition. In: ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume 2., IEEE (1992) 349–352 vol.2
5. Franco, H., Cohen, M., Morgan, N., Rumelhart, D., Abrash, V.: Context-dependent connectionist probability estimation in a hybrid hidden Markov model-neural net speech recognition system. Computer Speech & Language **8**(3) (July 1994) 211–222
6. Abad, A., Pellegrini, T., Trancoso, I., Neto, J.: Context dependent modelling approaches for hybrid speech recognizers. Interspeech 2010 (September) (2010) 2950–2953

[4] MAVIR corpus. http://www.lllf.uam.es/ESP/CorpusMavir.html

7. Kirchhoff, K., Fink, G.A., Sagerer, G.: Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication **37**(3-4) (July 2002) 303–319
8. Karen Livescu, O.c.: Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 jhu summer workshop
9. Frankel, J., Wester, M., King, S.: Articulatory feature recognition using dynamic Bayesian networks. Computer Speech & Language **21**(4) (October 2007) 620–640
10. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing **20**(1) (January 2012) 30–42
11. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE (December 2011) 24–29
12. : ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms (April 2000)
13. Oh, K.S., Jung, K.: GPU implementation of neural networks. Pattern Recognition **37**(6) (June 2004) 1311–1314
14. Strigl, D., Kofler, K., Podlipnig, S.: Performance and Scalability of GPU-Based Convolutional Neural Networks. In: 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing, IEEE (February 2010) 317–324
15. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing **13**(3) (May 2005) 345–354
16. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Interspeaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing **16**(5) (July 2008) 980–988
17. Ghahramani, Z., Hinton, G.: The EM algorithm for mixtures of factor analyzers. Technical Report 1, Dept. of Comp. Sci., Univ. of Toronto, Toronto (1996)
18. Vaquero, C., Ortega, A., Lleida, E.: Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (May 2011) 4532–4535
19. Rose, R.: Rapid Speaker Adaptation Using Clustered Maximum-Likelihood Linear Basis With Sparse Training Data. IEEE Transactions on Audio, Speech, and Language Processing **16**(3) (March 2008) 607–616
20. Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Pardo, J.M., Rubio, A.: Development of Spanish Corpora for Speech Research (Albayzín). In: Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods. (1991)
21. Moreno, A., Lindberg, B., Draxler, C.: Speechdat-car: A large speech database for automotive environments. LREC 2000 (April) (2000)
22. Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M.I., Lleida, E.: Improving dialogue systems in a home automation environment. In: 1st International ICST Conference on Ambient Media and Systems. (February 2010)
23. Heuvel, H.V.D., Choukri, K., Gollan, C.: TC-STAR: New language resources for ASR and SLT purposes. In: In Proceedings of the LREC 2006. Number March (2006) 2570–2573

Albayzin Speech Synthesis Evaluation
System Descriptions

Aholab Speech Synthesizer for Albayzin 2012 Speech Synthesis Evaluation

Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sánchez, Ibon Saratxaga.

Aholab Signal Processing Lab - University of the Basque Country, Bilbao, Spain
{inaki, derro, eva, inma, ion, ibon}@aholab.ehu.es

Abstract. This paper describes the Text-to-Speech (TTS) system presented by Aholab-EHU/UPV in the Albayzin 2012 Speech Synthesis (SP) evaluation campaign. Due to the characteristics of the provided corpus (multiple emotions and little data per emotion) a statistical synthesis approach was selected, training a single model with all the speech available. Various Pitch Detection Algorithms (PDAs) were employed in order to reduce gross errors and a more robust duration prediction was achieved by means of combining several machine learning methods. AhoCoder, the vocoder developed by Aholab Laboratory, was used to parameterize and reconstruct the speech signal with high quality.

Keywords: speech synthesis, statistical synthesis, emotions

1 Introduction

The Albayzin SS evaluation compares the performance of different TTS systems built with a common Spanish speech database. This year the database includes as a novelty a parallel corpus with four emotions (happiness, sadness, surprise and anger) plus neutral style. Once the training data is released, participants have several weeks to build the voice. Then, each group is asked to synthesize several hundred test texts that will be evaluated to determine the quality of the synthetic voices in terms of: naturalness, similarity to the original speaker, identification of the intended emotion and the perceived emotional intensity.

AhoTTS [1] is the synthesis platform for commercial and research purposes that Aholab Laboratory has been developing since 1995. It has a modular architecture, and written in C/C++ it is fully functional in both UNIX and Windows operating systems. Up to this date, synthetic voices for Basque, Spanish and English languages have been created. Contrary to last edition in which we submitted two systems (a statistical one and a hybrid system), this year we only present a statistical one. The reasons for this approach will be explained later on.

This paper is organized as follows. First, we describe the characteristics of the system. In Section 3 the voice building process is explained. The evaluation results are presented and discussed in Section 4. And finally, some conclusions are drawn in Section 5.

2 System Overview

This year Aholab presents a statistical TTS based on HTS [2]. As HTS does not perform any kind of linguistic analysis, the output of the first module of AhoTTS had to be translated into proper labels containing phonetic and linguistic information.

The reasons for not using the concatenative hybrid system [3] that yielded good results in Albayzin2010 [4] are several. The hybrid system showed good naturalness and consistency in part due to the characteristics of the recordings provided. (e.g. the low pitch variability of the male speaker and the availability of more than 2 hours of recordings in the same style). Thanks to those characteristics almost no pitch modification was necessary in the selected sequence of natural units, preserving, that way, their naturalness and reducing the distortion due to incorrect pitch marking. In the Albayzin 2012 data, those characteristics are missing because of the larger pitch variance of emotions as happiness and surprise. Besides, the amount of data per emotion is smaller than in the past edition, thus, increasing the probability of not finding proper natural units in the corpus and introducing distortions by having to modify their prosody. The possibility of looking for the optimal units in all the corpus available (i.e. including all the emotions) was also discarded due to some past experiments with similar data, in which the concatenation of units from different recording styles yielded unnatural sound most of the time.

Therefore, statistical synthesis was the chosen method by Aholab laboratory for Albayzin 2012 SS evaluation campaign. On the one hand it usually produces a more robust synthetic voice with little training data thanks to the good generalization of the models for unseen contexts. On the other, it offers greater flexibility [5] to combine all the available data from different styles, as long as it is properly labeled.

Our system was built using all the emotional data at once by simple adding a sentence level emotional label which indicated the emotion portrayed by the speaker during the recording session. With that procedure the statistical training algorithm was capable of combining data from different styles when they were similar enough, and otherwise modeling emotion specific characteristics.

2.1 Linguistic Processing

This first module performs several language dependent tasks. Text normalization and grapheme to phoneme conversion are conducted by means of rules, whereas POS tagging uses a specific lexicon and some simple disambiguation rules. The following features have been encoded into the context labels used by HTS. In short, there are the same features used in Albayzin 2010 removing the Intonation Break information and adding the emotion/style one:

- **Phoneme level:**
  - SAMPA label of the current phoneme.
  - Labels of 2 phonemes to the right and 2 phonemes to the left.
  - Position of the current phoneme in the current syllable (from the beginning and from the end).

- **Syllable level:**
  - Number of phonemes in current, previous and next syllables.
  - Accent in current, previous and next syllables.
  - Stress in current, previous and next syllables.
  - Position of the current syllable in the current word (from the beginning and from the end).
  - Position of the current syllable in the current accent group.
  - Position of the current syllable in the current sentence.
  - Position of the current syllable after the previous pause and before the next pause.
- **Word level:**
  - Simplified part-of-speech tag of the current, previous and next words (content/function).
  - Number of syllables of the current, previous and next words.
  - Position of the current word in the sentence (from the beginning and from the end).
  - Position of the current word after the previous pause and before the next pause.
- **Accent level:**
  - Type of current, previous and next accent groups, according to the accent position.
  - Number of syllables in current, previous and next accent groups.
  - Position of the current accent group in the sentence (from the beginning and from the end).
  - Position of the current accent group after the previous pause and before the next pause.
- **Pause context level:**
  - Type of previous and next pauses.
  - Number of pauses to the right and to the left.
- **Sentence level:**
  - Type of sentence.
  - Number of phonemes.
  - Number of syllables.
  - Number of words.
  - Number of accent groups.
  - Number of pauses.
  - Type of emotion/style.

2.2 Prosody Prediction

The intonation prediction is performed by the robust Multi-space Distribution MSD HMM [6], whereas the duration is predicted externally by a combination of CART and Random Forests (RF) [7] machine learning techniques. Durational data was divided into three broad classes (vowels, voiced consonants and unvoiced consonants) and optimal variables, such as the minimum occupancy per leaf or the number of trees (in RF), were determined by means of a ten-fold CV procedure. Table 1

shows the best results (i.e. using RF for vowels and CART for the other two classes) measured by the Pearson Correlation Coefficient.

Phoneme Class	Correlation (Train)	Correlation (Test)
Vowels	0.922	0.814
Voiced Consonants	0.833	0.784
Unvoiced Consonants	0.843	0.780

Table 1. Duration Prediction with 10-fold CV

We did not perform any kind of break insertion at synthesis time. So, only ortographic punctuation marks were considered as pauses.

2.3 Waveform Generation

AhoCoder [8], a Harmonic plus Noise Model (HNM) based vocoder was used in order to generate the synthetic speech from the estimated parameter sequence: spectrum, pitch and maximum voiced frequency (MVF).

3 Voice Building

Organizers provided a Spanish Emotional Voices (SEV) database recorded by GTH [9]. It contains data from the professional male speaker (Joaquín), simulating four full-blown emotions (happiness, sadness, anger and surprise) and a neutral speaking style, in a studio recording environment. The corpus for development contains more than three hours of acted-emotions recordings (approximately 40 minutes per style). In addition to the audio data and texts, ElectroGlottoGraph (EGG) waveform files were also available.

The training process of the statistical parametric voice is automatically done, once context labels and segmentation marks are ready and proper questions to build the trees are set.

3.1 Segmentation

Although the organizers provided segmentation labels, we decided to segment the whole corpus again with HTK toolkit [10] using the phoneme sequence provided by our own grapheme to phoneme converter. First, monophone models were trained from a plain start. Then, tied-state triphone models were trained and new labels obtained by means of forced alignment. During this first pass, the insertion of short pauses (SPs) was allowed at word boundaries. As an automatic post-processing before the second pass, pauses smaller than a threshold were eliminated and phoneme durational outliers were analyzed to decide whether to insert a pause next to them or not. Models were retrained after these changes and a second forced alignment was performed. This time no SPs were allowed. Finally, pause boundaries were automatically refined with a

simple processing based on phone duration and energy threshold. No manual revision of the segmentation labels was done.

3.2 Feature Extraction

All the language related features were extracted from our linguistic processing module. The extraction of the acoustic features consists of several steps. First, all the signals (speech and EGG) were down-sampled to 16kHz. Then, power normalization was performed by measuring the mean power in the middle of the vowels for all the sentences, and then normalizing each interpause interval. Afterwards, pitch contours were detected combining three different methods in order to avoid gross errors (our own PDA [11], get_f0 from Snack Toolkit and Praat). As far as the HTS training is concerned, the following parameters were extracted: f0 + 40 MFCCs + MVF.

Informal listening tests were done comparing the synthesis results for different pitch estimation inputs, and the best performance was obtained by extracting the pitch from the EGG signal with the three different PDAs and refining the combined pitch during the harmonic analysis of AhoCoder. The only exception to that rule was the surprise emotion for which the EGG recording was quite noisy. Therefore, the pitch was extracted from the speech signal for this emotion.

3.3 Impact of 'type of emotion' Label

As mentioned before, in order to use all the data to train the statistical models a new label was added indicating the type of emotion at sentence level. That way, we tried to overcome the little data per emotion available. To discover the importance of this new label we analyze the spectrum and pitch trees for all their five states. As far as pitch is concerned, the first appearance of the label in a question is before the third level in any of the trees. That shows that the tree leafs are rapidly clustered depending on their emotional content, especially for surprise and sadness emotions (which are the ones with highest and lowest variance respectively). As far as the spectrum is concerned, similar results are obtained. The question 'type of emotion' appears before the third level of the trees, and sadness and surprise are the most discriminating emotions (both had a distinctive voice quality).

Informal listening tests indicated that using all the available emotional data to build a single model produced a more consistent sound than building separate models per emotion. Being the sentences uttered by the same speaker, recorded in the same conditions and having each emotion correctly identified, it seems that the statistical modeling is able to make more robust models by clustering the similar frames of speech coming from different emotions.

4 Evaluation Results

Participants were asked to synthesize more than one hundred sentences for each emotion plus neutral style. Listeners performed 4 evaluation tasks: (i) Mean Opinion

Score (MOS) for the overall naturalness or quality, (ii) MOS for similarity to the original speaker, (iii) identification ability for the intended emotion and (iv) MOS for the perceived emotional intensity or strength.

All these four average metrics were normalized (into a range from 0 to 1) and a single measure called Emotional Performance measure (EP-measure) was defined in order to determine the best system. EP-measure is a combination of all the four metrics into one final score per system.

In this edition, 4 systems took part in the evaluation campaign. Each system was identified by a letter from B to E, being A reserved for the natural speech. Our synthetic system was identified by letter B and got the best EP-measure for neutral style, and the best average normalized speaker similarity. The results for each task are analyzed in the following subsections.

4.1 Naturalness or Speech Quality

In this task the overall quality of the speech was evaluated in a 5-point scale ranging from 1 ("very bad") to 5 ("very good)". This scale was then linearly transformed to a [0.0 - 1.0] range. Our system got an average 0.44 score, ranging from 0.49 for anger to 0.39 for sadness. The slightly worse results for sadness could be explained by the difficulty to properly extract the pitch and the voiceness for this emotion due to the special voice characteristic that it usually has.

4.2 Similarity to the Original Speaker

Listeners had to rate how similar synthetic or natural speech was when compared to the neutral voice of the speaker, in a 5-point scale ranging from 1 ("Sounds like a totally different person") to 5 ("Sounds like exactly the same person"). Our system got a median value of 3 points for all the emotions but for sadness, for which 2 points were scored. As far as the normalized speaker similarity is concerned, our system got the best results with an average of 0.46, ranging from 0.39 for sadness to 0.52 for neutral and happiness. Being our system a statistical one, we are a bit surprise by the fact that we got the best average results. That might mean that the rest of the systems are not concatenative ones and that they might have adapted their voices from an average model built with external data.

4.3 Intended Emotion

Listeners had to identify the intended emotion from a limited set of emotional categories: happiness, anger, surprise, sadness, neutral or another. Table 2 shows the confusion matrix for system B. All the 4 emotions plus neutral style were identified with a rate much greater than chance (0.166, taking into account "other"). That rate ranges from 0.35 for surprise to a far high 0.83 for neutral style. Sadness has also a good identification rate in spite of getting the worse results in the previous tasks. Besides, most emotions are confused with neutral style. The average emotion identification rate is 0.53.

Intended Emotions	Identified Emotions					
	Happiness	*Anger*	*Surprise*	*Sadness*	*Neutral*	*Other*
Happiness	**0.38**	0.12	0.19	0.06	0.21	0.04
Anger	0.02	**0.48**	0.06	0.10	0.34	-
Surprise	0.16	0.06	**0.35**	0.18	0.18	0.08
Sadness	0.04	0.02	0.02	**0.56**	0.32	0.04
Neutral	0.04	0.02	-	0.10	**0.83**	0.02

Table 2. Confusion matrix for our system

4.4 Perceived Emotional Intensity

Listener had to assess the emotional strength or intensity in a 5-point scale from "very weak" to "very strong". The average strength score of our system is 0.41, ranging from 0.33 for neutral to 0.47 for happiness and sadness. It is somehow unexpected that being neutral the "emotion" with highest identification rate, it gets the worse emotional intensity. Maybe listeners were confused when having to evaluate the emotional intensity of a neutral or non emotional speech.

5 Conclusions

This has been our third participation in the Albayzin SS evaluation campaign. We built a statistical synthetic voice, using all the available emotional data and including a context label in order to identify each emotion at sentence level.

4 systems took part in the evaluation and ours got the best EP-measure for neutral style the best average normalized speaker similarity. All the emotions were identified far above chance levels and we could state that our approach training a single voice with all the available emotional data has succeeded. As in previous editions, there is still a big gap between the best synthetic system and natural speech, but that gap seems to be even larger when dealing with emotional data.

6 Acknowledgements

The authors would like to thank the organizers of Albayzin SS 2012 and the developers of all the tools employed during the voice building process.

This work has been partially supported by UPV/EHU (Unidad de Formación e Investigación UFI11/30), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Saiotek Project, PE11UN081).

7 References

1. Hernaez, I., Navas, E., Murugarren, J.L., and Etxebarria, B. Description of the AhoTTS conversion system for the Basque language. In: Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis. (2001)
2. "HMM-based Speech Synthesis System (HTS)," http://hts.sp.nitech.ac.jp/
3. Sainz, I., Erro, D., Navas, E., Hernáez, I., Sánchez, J., Saratxaga, I., Odriozola, I., and Luengo, I. Aholab Speech Synthesizers for Albayzin2010. In: Fala2010. pp. 343--348 (2010)
4. Méndez, F., Docío, L., Arza, M., and Campillo, F. The Albayzín 2010 Text-to-Speech Evaluation. In: Fala2010. pp. 317--340 (2010)
5. Zen, H., Tokuda, K., and Black, A.W. Statistical parametric speech synthesis. In: Speech Communication. vol. 51. no. 11. pp. 1039--1064 (2009)
6. Tokuda, K., Masuko, T., Miyazaki, T. Multi-space probability distribution HMM. In: IEICE Trans. Inf. & Syst. vol. E85-D. no. 3. pp. 455--464 (2002)
7. Breiman, L. Random forests. In: Machine learning. vol. 25. no. 2. pp. 5--32 (2001)
8. Erro, D., Sainz, I., Navas, E., and Hernáez, I. HNM-Based MFCC+f0 Extractor Applied to Statistical Speech Synthesis. In: ICASSP 2011. pp. 4728--4731 (2011)
9. Barra-Chicote, R., Montero, J.M., Macias-Guarasa, J., Lutfi, S., Lucas, J.M., Fernández, F., D'Haro, L.F., San-Segundo, R., Ferreiros, J. , Córdoba, R., Pardo, J.M. Spanish Expressive Voices: corpus for emotion research in Spanish. In: Proceedings of the 6th conference of Language Resources & Evaluation (Workshop on Corpora for Research on Emotion and Affect). pp 60--70. Marrakech, Morocco (2008)
10. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Olla-son, D., Povey, D., Valtchev, V., and Woodland, P.C. The HTK Book. version 3.4 (2006)
11. Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sánchez, J., and Sainz, I. Evaluation of Pitch Detection Algorithms Under Real Conditions. In: IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'07. pp 1057--1060 (2007)

Simple4All proposals for the Albayzin Evaluations in Speech Synthesis

Jaime Lorenzo-Trueba[1], Oliver Watts[2], Roberto Barra-Chicote[1], Junichi Yamagishi[2], Simon King[2], and Juan M. Montero[1]

[1]Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain
[2]CSTR, University of Edinburgh, United Kingdom
jaime.lorenzo@die.upm.es

Abstract. Simple4All is a European funded project that aims to streamline the production of multilanguage expressive synthetic voices by means of unsupervised data extraction techniques, allowing the automatic process of freely available data into flexible task specific voices. In this paper we describe three different approaches for this task, the first two covering enhancements in expressivity and flexibility with the final one focusing on the development of unsupervised voices. The first technique introduces the principle of speaker adaptation from average models consisting of multiple voices, with the second being an extension of this adaptation concept into allowing the control of the expressive strength of the synthetic voice. Finally, an unsupervised approach to synthesis capable of learning from unlabelled text data is introduced in detail.

Keywords: Emotional Speech Synthesis, Unsupervised Synthesis, Emotional Strength Control

1 Introduction

One of the goals in the Simple4all[1] project is the automatic modification of the speaking style of a neutral or expressive voice without requiring the recording of additional data. It should also be able to maintain the synthesized voice quality of the source voice but at the same time providing high recognition rates of the target expressive style. Additionally, because expressivity in real life is not a discrete space but a continuous space it is also required for the system to be able to mimic this property. Another aim of the project is to be able to generate all these voices with unlabelled data, as that significantly increases the potential sources of training data from which to produce the models.

This paper proposes three different systems that aim to fulfil different aspects of the project. The first and second systems (systems A and B) tackle the problem of enhancing and modifying the expressivity of voice models with as little data as possible by making use of adaptation techniques. The third system

[1] www.simple4all.org/

(system C) provides a state of the art system capable of training unsupervised synthetic voices.

Albayzin 2012, then, proves to be the perfect challenge for testing our systems as its main focus on enhancing the naturalness and emotional strength and identification rates while keeping the similarity with the source's natural voice is the definite proving ground for the prosed systems.

Regarding the structure of the present paper, Sections 2, 3 and 4 describe the three proposed systems, namely: Average Model based Voices, Adaptation-based Emotional Strength Controlled Voices and the Unsupervised Front-End in that order. Section 3 also includes a short review on a previous perceptual analysis of the system to justify its inclusion. Finally, section 5 includes the results of the challenge and conclusions to be drawn from them.

2 System C: Average Model based Voices

Because sometimes there is only a small amount of data available for training the voices, or even because the processing time of the training is critical, systems based in the creation of average models and obtaining adapted voices from them such as System **C** have been created. This particular system is based on obtaining the acoustic models of the different emotions of the speakers through a model-space SAT algorithm [1] to then proceed to use a shared decision-tree based clustering algorithm such as [2] in order to tie the parameters and define the speaker average model. From this average voice that encompasses all the different versions of the speaker voice (neutral, happy, sad, angry and surprised in this evaluation), the particular emotional model is then adapted using the CSMAPLR algorithm [3] (Figure 1).

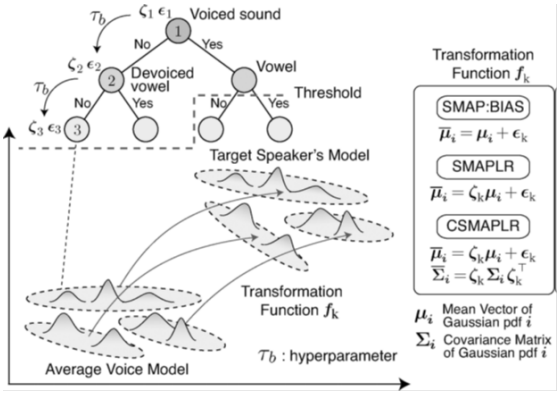


Fig. 1. Schematic defining the CSMAPLR average voice generation process.

The voices obtained in this fashion are much more robust in situations in which the training data is limited, as it pools all the available resources in order

to attain stability, and then modifies the parameters so that the obtained model resembles more closely the intended voice.

3 System D: Adaptation-based Emotional Strength Controlled Voices

It is known that HMM-based modelling introduces a smoothing in the synthetic voices, reducing the expressive capabilities of the model. This side effect is further enhanced by the adaptation process. Consequently, in this particular task where it is important to synthesize expressive voices it becomes necessary to find a way to enhance said expressivity. With that purpose in mind we developed system **D**. In this system, the adapted models of the different emotions are obtained from the average voice in the same fashion as in system **C**, but they are only used to obtain the transformation function capable of morphing the neutral model into the different emotional models (Figure 2). Then, through a control ratio it becomes possible to either enhance or attenuate the expressivity of the transformed model, allowing for a continuous modeling of the emotional space.

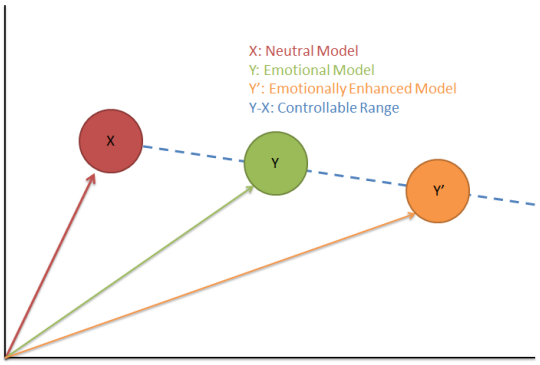


Fig. 2. Representation of the emotional strength control system.

3.1 Analysis of the ES-Control effects

This system has been studied through a perceptual test, which analysed the emotion identification rates, perceived naturalness and perceived emotional strength for the adaptation-based system with control ratios ranging from 0.0 to 2.0, always comparing it with its natural voice.

Regarding identification rates, for control ratios of 1.5 times the plain adapted emotion and upwards the results become comparable to those of natural voice, sometimes even surpassing them. This trend has been proven to be stable at

least up to control ratios of 2.0, but it would be hard to extrapolate them further than that due to natural speech rarely being so full blown emotional. For naturalness, considering the natural voice obtained a 4 out of 5 in the likert scale, the ES-controlled voice manages to attain constant values of 3, stable in all the considered ratios, which proves the usefulness of enhancing the expressivity in this fashion as it does not alter the attained speech quality. Finally, the perceived emotional strength was seen to be practically linear throughout the examined control ratios range, corroborating the hypothesis that a continuous emotional strength space can be attained. Additionally, it was also since control ratios between 1.5 and 2.0 that the perceived strength results matched those of natural voice, once again sometimes surpassing them.

All in all, the conclusion of this study was that applying emotional strength control through adaptation can provide significant enhancements to the perceivable expressivity of the synthetic voices without incurring in naturalness and speech quality penalties for control ratios of at least up to 2.0, with possibly 1.5 being the optimal value that does not suffer from over-exaggeration effects in the synthesized voice.

4 System E: Unsupervised Front-End

System **E** was built to test the prototype text-processing modules which are being developed as part of the *Simple4All* project. The aim of developing these modules is to provide a TTS front-end which makes few implicit assumptions about the target language, and which can be configured with minimal effort and expert knowledge to suit arbitrary new target languages. To this end, the modules rely on resources which are intended to be universal, such as the Unicode character database[2], and make use of unsupervised learning so that unlabelled text resources can be exploited without the need for costly annotation. The initial version of this front-end is based very closely on the ideas outlined in [4], and is a re-implementation with some modifications of the system that was there used to build synthetic voices in English, Romanian and Finnish. The version of the front-end used in the current Challenge will be briefly described here, as well as the training of acoustic models using the annotation provided by it.

4.1 Text Analysis

Data The ISO 8859-1 text of the transcriptions provided for the Challenge was first manually converted to UTF-8 encoding, which is required by the front-end. After this initial conversion, however, text processing is fully automatic as described below. In addition to the text transcription of the speech, we used an additional 1.4 million words of untranscribed Spanish text. This text was selected to roughly match the domain of the speech corpus: the entire text of Cervantes' *Don Quijote* (c. 400,000 words) was taken from Project Gutenberg[3] to match

[2] `www.unicode.org/ucd/`

[3] `www.gutenberg.org`

the Quijote portion of the corpus, and the text of c. 1800 news stories drawn randomly from those published by the *El Mundo* newspaper in 2006 (giving c. 1 million words) was used to match the remaining part of the corpus

Tokenisation All language-specific knowledge used to provide the chunking of text into words and punctuation symbols is derived from the Unicode character database. Each input character is assigned a *coarse category* by lookup in the that database. The three coarse categories used are formed by merging Unicode general categories: coarse category *letter* is made by made by combining general categories L and N, coarse category *space* maps exactly to general category Z, and all remaining general categories map to a coarse category called *punctuation*. Tokenisation is performed by placing token delimiters at the ends of contiguous sequences of characters belonging to the coarse category *letter*. The chunks between these delimiters are tokens. For example (if whitespace is represented for clarity by an empty box) the training utterance:

sí.□con□seguridad.

is chunked into the following six tokens:

sí .□ con □ seguridad .

Furthermore, the coarse categories are used to assign each token a token class, from an inventory of three classes that share the names of the coarse categories. The coarse categories are placed in the following order of precedence: letter, punctuation, space. A given token is classified by traversing this sequence of categories from left to right; if all characters in that token belong to the current category or to a category on the left, the token is assigned to that token class. The above chunks are assigned the following token classes:

letter punct. letter space letter punct.

For implementational reasons (e.g. to enable use of tools which can only handle ASCII characters and to avoid confusion with special symbols such as field delimiters during processing), each input character is assigned a *safetext* form, consisting of a string of characters belonging to the 52 upper- and lower-case letters of the English alphabet. Safetexts consisting of multiple characters but corresponding to a single character of surface text are delimited by underscores to allow unambiguous mapping back to surface forms as needed. Where a surface form is not already a safetext, one is constructed automatically from the name of the character in the Unicode database. The 6 chunks above are rewritten in safetext form as follows:

s_LATINSMALLLETTERIWITHACUTE_
_FULLSTOP__SPACE_
con
SPACE
seguridad
FULLSTOP

All characters encountered in the text of the training data are stored in a text file along with their automatically generated coarse categories and safetexts. This allows a user to manually intervene to correct bad categorisation of characters due to text encoding mistakes or due to errors in the Unicode database, and to specify more user-friendly safetexts than the automatically generated ones (see e.g. the long-winded *_LATINSMALLLETTERIWITHACUTE_* for the character *í* in the example above). For building the voices presented here, however, automatically generated categories and safetexts were used with no manual intervention.

Note that the present front-end requires text to be expanded fully before it is input to the system, and does not handle numerals and abbreviations. Correctly handling such non-standard words in a way that requires only minimal expert supervision is a topic of ongoing research [5].

Naive 'lexicon' and time alignment A naive 'lexicon' is used, in which the safetexts of letters of 'letter'-class tokens are used directly as the names of speech modelling units, in place of the phonemes of a conventional front-end. This has given good results for languages with transparent alphabetic orthographies such as Romanian, Spanish and Finnish, and can give acceptable results even for languages with less transparent orthographies, such as English [4, 6–8]. Using the pronunciations provided by this lexicon, a set of labels is initialised for the speech part of the database by iteratively estimating a set of HMMs and using these to force-align the speech with the labels using a procedure based very closely on that described in [9]. Tokens which are assigned by the tokeniser to the *punctuation* and *space* token classes are allowed by the naive lexicon to be pronounced both as a silence symbol (*sil*) or as a non-emitting symbol (*skip*). As well as determining the timing of letter-boundaries, therefore, the forced alignment procedure determines which space and punctuation tokens are realised as a pause, and which are skipped.

Letter- and word-representations The system makes use of no expert-specified categories of letter and word, such as phonetic categories (vowel, nasal, approximant, etc.) and part of speech categories (noun, verb, adjective, etc.). Instead, features that are designed to stand in for such expert knowledge but which are derived fully automatically from the distributional analysis of the text corpus are used. The distributional analysis is conducted via vector space models (VSMs); the VSM was applied in its original formulation as a model for Information Retrieval to the characterisation of documents. VSMs are applied to TTS in [4], where models are built at various levels of analysis (letter, word and utterance) from large bodies of unlabelled text. To build these models, co-occurrence statistics are gathered in matrix form to produce high-dimensional representations of the distributional behaviour of e.g. word and letter types in the corpus. Lower-dimensional representations are obtained by approximately factorising the matrix of raw co-occurrence counts by the application of slim singular value decomposition. This distributional analysis places textual objects

in a continuous-valued space, which is then partitioned during the training of TTS system components such as acoustic models for synthesis or decision trees for pause prediction. For the present voices, a VSM of letters was constructed by producing a matrix of counts of immediate left and right co-occurrences of each letter type, and from this matrix a 5-dimensional space was produced to characterise letters. Token co-occurrence was counted with the nearest left and right neighbour tokens which are *not* of the class *space*; co-occurrence was counted with the most frequent 250 tokens in the corpus. A 10-dimensional space was produced to characterise tokens.[4]

Pause prediction The system uses a decision tree to predict whether a token of class *space* or *punctuation* is realised as a pause or not. Data for training the tree is produced from the time-aligned transcriptions of the training data. The predictor variables used for tree training are the token class of the token in question (i.e. whether it is punctuation or space) and the VSM features of the tokens preceding and following the token. The annotation of training data is done by detection of silence in the audio during forced alignment as already described. At run-time, the tree's predictions are used.

Rich contexts Information extracted from the utterance structures resulting from text processing is used to create a set of rich contexts for each speech unit in the database. Features include the identity of the letter to be modelled and those of its neighbours (within a 5-letter window), the VSM values of each of those letters, and the distance from and until a word boundary, pause, and utterance boundary. Word VSM features were not included directly in the contexts, but were used by the decision tree for predicting pauses at runtime. It should be emphasised that there is nothing language-specific about these features: they are generally applicable to any language making use of an alphabetic script and marking word boundaries orthographically.

4.2 Acoustic Models

Emotion-dependent acoustic models were built for each of the angry, happy, neutral and surprised subsets of the database, using the distributed 48kH waveforms and the acoustic parameters described in [11], and using a now standard speaker dependent recipe described in [12]. The *sad* subset of the data created problems for the extraction of STRAIGHT features, and so the least problematic part of that subset (the *Quijote* section) was used to adapt the acoustic models built for the neutral condition to the sad condition (using a combination of the CSMAPLR adaptation and MAP adaptation techniques, as in [13]), although it was found that using the unadapted neutral duration model gave best results. The waveforms of the synthesised test-set were downsampled to the required 16kH for submission.

[4] The package *Gensim* [10] was used for performing the singular value decomposition needed to obtain these features.

5 Results

The first conclusion that can be extracted is that even if synthetic voices do not present the same quality results as the natural voice yet, although for emotion identification rates, emotional strength and speaker similarity the obtained results show that the presented systems can compare and sometimes even surpass the recognition potential of the natural voice. This is a feat that encourages us for developing even further the systems in order to one day obtain the most natural possible voices. In a system-by-system analysis of the results, system **C**,

SYSTEM	Speech Quality	Emotional Strength	Speaker Similarity	Emotion Identification Rate	Relative Performance
A	0.87	0.71	0.43	0.78	**1.00**
B	0.44	0.41	0.46	0.53	**0.32**
C	0.57	0.53	0.42	0.65	**0.51**
anger	0.62	0.64	0.37	0.78	0.51
happiness	0.57	0.61	0.48	0.65	0.80
neutral	0.51	0.33	0.47	0.61	0.60
sadness	0.49	0.52	0.41	0.65	0.48
surprise	0.65	0.57	0.38	0.50	0.29
D	0.53	0.54	0.39	0.63	**0.46**
anger	0.61	0.67	0.36	0.82	0.53
happiness	0.53	0.59	0.40	0.34	0.38
neutral	0.55	0.31	0.36	0.76	0.58
sadness	0.49	0.56	0.32	0.79	0.47
surprise	0.51	0.59	0.51	0.37	0.25
E	0.41	0.43	0.42	0.46	**0.27**
anger	0.35	0.38	0.46	0.30	0.13
happiness	0.44	0.53	0.35	0.46	0.36
neutral	0.43	0.32	0.34	0.74	0.46
sadness	0.43	0.43	0.43	0.38	0.26
surprise	0.42	0.51	0.55	0.38	0.21

Table 1. Normalized results for the proposed systems [14]. All four measures take values from 0 to 1, with the performance being obtained as $4 * (SQ * ES * SS * EIR)/(SQ + ES + SS + EIR)$ and then normalized by the natural speech (system **A**).

is clearly the one with the best overall results, constantly placing first or second in all the categories excepting speaker similarity. This results prove the usefulness of using averaged data and then adapting into the particular emotion in this situation in which the training data is not very extensive. One of the best conclusions that can be extracted from the results of this system is that they are significantly stable, not showing any kind of unexplainable dip in any emotion or measured category. The only exception would be the recognition rate of surprise, which can be justified by the significant confusion it presented with happiness.

System **D** shows how the strength control process is promising, although it can be improved. The expected results would be to obtain the better identification rates and emotional strength at the cost of speech quality, and so it is but with some significant constraints. First of all, the confusion between happiness and surprise was greatly enhanced to the point where both emotions become almost unrecognisable between each other. This is thought to be because all the different features were extrapolated, and it is expected that extrapolating only some particular features (such as modifying the spectral parameters while leaving the F0 intact) would help solve this problem. Emotional strength results show that they are marginally higher or lower than those of system **C**, but not in a statistically relevant way. It is also expected that applying the partial control would help relief this perception problems.

Finally, the unsupervised system (system **E**), shows some very promising results in the sense that given the handicap of being unsupervised, all the different measures are comparable with other supervised systems in the competition, sometimes even surpassing it in speaker similarity. The biggest problem is presented when considering the emotional measures was that the voices obtained through this system, while very clear and natural sounding, are sometimes too flat-sounding and neutral. One additional cause for this was that the feature extraction section of this system had some problems when extracting the fundamental frequency of the training data, which made the resulting voiced even more flat. Nevertheless the synthetic voice still keeps its natural properties as it can be seen in the speaker similarity measures, in which this system sometimes obtains the best results.

6 Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO(TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo has been funded by Universidad Politecnica de Madrid under grant SBUPM-QTKTZHB. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

References

1. J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.
2. J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–1233.

3. J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.
4. Oliver Watts, *Unsupervised Learning for Text-to-Speech Synthesis*, Ph.D. thesis, University of Edinburgh, 2012.
5. Ruben San-Segundo, Juan M. Montero, Veronica Lopez-Ludeña, and Simon King, "Detecting acronyms from capital letter sequences in Spanish," in *Proc. Interspeech*, Portland, Oregon, USA, Sept. 2012.
6. A. Black and A. Font Llitjos, "Unit selection without a phoneme set," in *IEEE TTS Workshop 2002*, 2002.
7. G.K. Anumanchipalli, K. Prahallad, and A.W. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 4645–4648.
8. Matthew P. Aylett, Simon King, and Junichi Yamagishi, "Speech synthesis without a phone inventory," in *Interspeech*, 2009, pp. 2087–2090.
9. Robert A. J. Clark, Korin Richmond, and Simon King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.
10. Radim Rehurek and Petr Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, May 2010, pp. 45–50, ELRA.
11. Junichi Yamagishi and Oliver Watts, "The CSTR EMIME HTS System for Blizzard Challenge," in *Proc. Blizzard Challenge 2010*, Sept. 2010.
12. H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.
13. Junichi Yamagishi, Heiga Zen, Tomoki Toda, and Keiichi Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge 2007*, aug 2007.
14. Iberspeech2012, "2012 albayzin evaluations: speech-synthesis results," 2012.